(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 April 2001 (05.04.2001)

PCT

(10) International Publication Number
WO 01/23401 A2

(51) International Patent Classification⁷: C07H 21/00

(21) International Application Number: PCT/US00/26708

(22) International Filing Date:
28 September 2000 (28.09.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/408,393 28 September 1999 (28.09.1999) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 09/408,393 (CIP)
Filed on 28 September 1999 (28.09.1999)

(71) Applicant *(for all designated States except US)*: MAXYGEN, INC. [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: WELCH, Mark [US/US]; 25 Montalban Drive, Fremont, CA 94536 (US). NESS, Jon, E. [US/US]; 1220 North Fairoaks Avenue #2115, Sunnyvale, CA (US). GUSTAFSSON, Claes [SE/US]; 1813 Bayview Avenue, Belmont, CA 94002 (US). STEMMER, Willem, P., C. [NL/US]; 108 Kathy Court, Los Gatos, CA 95030 (US). MINSHULL, Jeremy [GB/US]; 842 Hermosa Way, Menlo Park, CA 94025 (US).

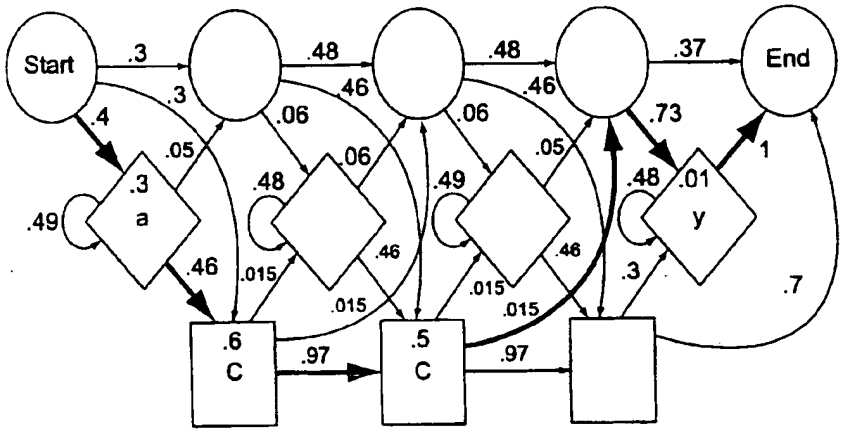
(74) Agents: QUINE, Jonathan, Alan; The Law Offices of Jonathan Alan Quine, P.O. Box 458, Alameda, CA 94501 et al. (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: USE OF CODON-VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SEQUENCE RECOMBINATION



(57) Abstract: Methods of providing recombination libraries that include codon-varied oligonucleotide sequences are described. Codon-varied oligonucleotides are synthesized using trinucleotide or mononucleotide phosphoramidite sequences, and are derived from homologous or non-homologous nucleic acid sequences, or combinations of such sequences. Various methods of recombining codon-varied oligonucleotide sequences to expedite artificial evolution are also described. The present invention additionally relates to various integrated systems that are optionally used to automate these recombination methods.

WO 01/23401 A2

Published:

— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Use of Codon-Varied Oligonucleotide Synthesis for Synthetic Sequence Recombination

CROSS-REFERENCES TO RELATED APPLICATIONS

This application is a continuation-in-part of U.S. Patent Application No. 09/408,393, filed on September 28, 1999, Welch et al., entitled USE OF CODON-VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING, the disclosure of which is incorporated by reference. The present application claims priority to and the benefit of this related application, pursuant to 35 U.S.C. 119, 35 U.S.C. 120 and any other applicable statute or rule.

COPYRIGHT NOTIFICATION

FIELD OF THE INVENTION

The present invention relates to methods of providing recombinant libraries that include codon-varied oligonucleotide sequences. Codon-varied oligonucleotides can be synthesized using trinucleotide or mononucleotide phosphoramidite sequences, and can be derived from homologous or non-homologous nucleic acid sequences, or combinations of such sequences. In turn, codon-varied oligonucleotide sequences can be utilized to generate diversity in various methods of recombination, mutation, artificial evolution, or the like.

BACKGROUND OF THE INVENTION

The use of trinucleotide phosphoramidites in solid-phase DNA synthesis was previously thought to be unfeasible, as only marginal yields could be achieved. Sondek, J. and Shortle, D. (1992) *J. Immunol.*, 149, 3903-3913. These poor results were attributed to the steric bulk of the trinucleotide molecules. *Id.* However, it has since been shown that trinucleotide phosphoramidites representing

codons for all 20 amino acids can be successfully used to introduce entire codons into oligonucleotides in automated, solid-phase DNA synthesis and thus can function as excellent reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Virnekäs, B., *et al.*, (1994) *Nucleic Acids Res.*, 22, 5600-5607. Other references involving the synthesis of trinucleotide phoshoramidites, their subsequent use in oligonucleotide synthesis, and related issues are described in, e.g., Kayushin, A. L. *et al.*, (1996) *Nucleic Acids Res.*, 24, 3748-3755, Huse, U.S. Pat. No. 5,264,563 "PROCESS FOR SYNTHESIZING OLIGONUCLEOTIDES WITH RANDOM CODONS," Lyttle *et al.*, U.S. Pat. No. 5,717,085 "PROCESS FOR PREPARING CODON AMIDITES," Shortle *et al.*, U.S. Pat. No. 5,869,644 "SYNTHESIS OF DIVERSE AND USEFUL COLLECTIONS OF OLIGONUCLEOTIDES," Greyson, U.S. Pat. No. 5,789,577 "METHOD FOR THE CONTROLLED SYNTHESIS OF POLYNUCLEOTIDE MIXTURES WHICH ENCODE DESIRED MIXTURES OF PEPTIDES," and Huse, WO 92/06176 "SURFACE EXPRESSION LIBRARIES OF RANDOMIZED PEPTIDES."

The inventors and their co-workers have developed various rapid artificial evolution techniques for creating improved industrial, agricultural, and therapeutic genes and encoded proteins including via oligonucleotide-mediated recombination or "family" shuffling. *See, e.g.*, "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed September 28, 1999 (USSN 09/408,392) and "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed January 18, 2000 (PCT/US00/01203), both of which are incorporated by reference herein in their entirety for all purposes. Similarly, *in silico* recombination methods utilizing oligonucleotide substrates were described in, e.g., "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed January 18, 2000, (PCT/US00/01202) and "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed July 18, 2000 (USSN 09/618,579), both of which are incorporated by reference herein in their entirety for all purposes.

Additional oligonucleotide-mediated recombination methods would be desirable. The present invention provides codon-based oligonucleotide-mediated

recombination methods and related compositions, as well as a variety of additional features which will become apparent upon review of the following description.

SUMMARY OF THE INVENTION

The present invention provides recombination methodologies in which codon-varied oligonucleotides are recombined to provide recombined nucleic acid populations. Codon-varied oligonucleotides are synthesized, e.g., utilizing codon- or trinucleotide-based phosphoramidite coupling chemistry. This approach affords extensive flexibility to recombination processes, as codon-varied oligonucleotides can be based upon homologous or non-homologous nucleotide sequences, or even combinations of such sequences.

In a first aspect, the present invention is directed to a method of recombining codon-varied oligonucleotides to provide a population of recombined nucleic acids. It includes providing, hybridizing, and elongating a set of nucleic acids which include a plurality of codon-varied oligonucleotides (e.g., overlapping codon-varied oligonucleotides). For example, the codon-varied oligonucleotides are optionally produced by trinucleotide synthesis (e.g., automated trinucleotide synthesis, split-pool trinucleotide synthesis, or the like). In one embodiment, this method can include selecting at least first and second nucleic acids to be recombined, where the set of nucleic acid fragments includes a plurality of codon-varied nucleic acids which correspond to the first and second nucleic acids. The first and second nucleic acids can be homologous or non-homologous.

In one embodiment, the providing step of this method includes providing a solid-phase substrate sequence having a 5' terminus and at least one base, both of which have protecting groups thereon. The 5' protecting group of the solid-phase substrate sequence is then removed to provide a 5' deprotected solid-phase substrate sequence, which is then coupled with a selected trinucleotide phosphoramidite sequence. The trinucleotide has a 3' terminus, a 5' terminus, and three bases, each of which has protecting groups thereon. The coupling step yields an extended oligonucleotide sequence. Thereafter, the removing and coupling steps are optionally repeated. When these steps are repeated, the extended oligonucleotide sequence yielded by each repeated coupling step becomes the solid-phase substrate sequence of the next repeated removing step until a desired codon-varied oligonucleotide is obtained. This trinucleotide synthesis format can optionally

include coupling together one or more of: mononucleotides, trinucleotide phosphoramidite sequences, or oligonucleotides.

The providing step is optionally a "split-pool" synthesis format that includes providing solid-phase substrate sequences, each having a 5' terminus and at least one base, both of which have protecting groups thereon. The 5' protecting groups of the solid-phase substrate sequences are removed to provide 5' deprotected solid-phase substrate sequences, which are then coupled with selected trinucleotide phosphoramidite sequences. Each trinucleotide has a 3' terminus, a 5' terminus, and three bases, all of which have protecting groups thereon. The coupling step yields extended oligonucleotide sequences. Thereafter, the removing and coupling steps are optionally repeated. When these steps are repeated, the extended oligonucleotide sequences yielded by each repeated coupling step become the solid-phase substrate sequences of the next repeated removing step until extended intermediate oligonucleotide sequences are produced. In alternative embodiments, non-solid-phase synthesis methods are used.

Additional steps of the split-pool format optionally include splitting the extended intermediate oligonucleotide sequences into two or more separate pools. After this is done, the 5' protecting groups of the extended intermediate oligonucleotide sequences are removed to provide 5' deprotected extended intermediate oligonucleotide sequences in the two or more separate pools. Following this, these 5' deprotected intermediates are coupled with one or more selected mononucleotides, trinucleotide phosphoramidite sequences, or oligonucleotides in the two or more separate pools to yield further extended intermediate oligonucleotide sequences. In turn, these further extended sequences are pooled into a single pool. Thereafter, the steps beginning with the removal of the 5' protecting groups of the solid-phase substrate sequences to provide 5' deprotected solid-phase substrate sequences are optionally repeated. When these steps are repeated, the further extended oligonucleotide sequences, yielded by each repeated coupling step that generates those specific sequences, become the solid-phase substrate sequences of the next repeated removing step that includes those specific sequences, until desired codon-varied oligonucleotides are obtained.

Both synthetic protocols can optionally be performed in an automated synthesizer that automatically performs the steps. This aspect includes inputting

character string information into the automated synthesizer corresponding to the desired codon-varied oligonucleotides to be obtained, e.g., information corresponding to two or more nucleic acids to be recombined. Additionally, the protected solid-phase substrate sequences of both synthetic formats typically include 3' ends that are covalently attached to a solid support.

The hybridization step of the method described herein can occur *in vitro* or *in vivo*. The elongation step of this method optionally includes elongating the set of hybridized nucleic acid fragments with a polymerase (e.g., a thermostable polymerase), a ligase (e.g., a thermostable ligase), or both.

In one embodiment, the method of recombining codon-varied oligonucleotides optionally includes denaturing the population of recombined nucleic acids to provide denatured recombined nucleic acids. These denatured nucleic acids are then re-hybridized and in turn, elongated. In another embodiment of this method, the denaturing, re-hybridizing, and elongating steps are repeated at least once and optionally twice, three times, four times, or more. Finally, the resulting elongated re-hybridized recombined nucleic acids, from either embodiment, are selected for at least one desired trait or property.

In an additional embodiment of the method in which the denaturing, re-hybridizing, and elongating steps are repeated at least once, a plurality of members of the population of recombined nucleic acids is optionally selected for a desired trait or property to provide first round selected nucleic acids. This method optionally includes hybridizing an additional set of nucleic acid fragments to provide a population of further recombined nucleic acids. This method also optionally includes sequencing the first round selected nucleic acids, where the additional set of nucleic acid fragments is derived from the first round selected nucleic acids by aligning sequences of the first round selected nucleic acids to identify regions of identity and regions of diversity. The additional set of nucleic acid fragments is then synthesized to include a plurality of codon-varied oligonucleotides, each of which include subsequences corresponding to at least one region of diversity. The first round selected nucleic acids encode, e.g., polypeptides of about 50 amino acids or less, or larger peptides, e.g., about 60, about 70, about 80, about 90 amino acids or more. Furthermore, the additional set of nucleic acid fragments optionally include a plurality

of oligonucleotide member types which correspond to consensus region subsequences derived from a plurality of the first round selected nucleic acids.

In another aspect, the method of recombining codon-varied oligonucleotides optionally includes selecting at least one member of the population of recombined nucleic acids for at least one desired trait or property. Also, the set of nucleic acid fragments optionally includes a plurality of oligonucleotide member types that include consensus region subsequences derived from a plurality of homologous target nucleic acids. Further, the set of nucleic acid fragments, including a plurality of oligonucleotide member types, includes, alternatively, at least about 3, about 5, about 10, about 100, about 1,000 or more member types. Finally, the set of nucleic acid fragments optionally includes a plurality of homologous oligonucleotide member types that are present in either approximately equimolar amounts or approximately non-equimolar amounts.

In a second aspect, the invention provides a method of recombining at least two parental nucleic acids to provide at least one recombinant nucleic acid. This method includes providing a composition comprising at least one set of fragmented parental nucleic acids corresponding to the at least two parental nucleic acids. The set of fragmented parental nucleic acids includes a plurality of codon-varied oligonucleotides (e.g., overlapping codon-varied oligonucleotides). Next, the composition is hybridized to provide at least one hybridized nucleic acid. The at least one hybridized nucleic acid is then elongated to provide at least one recombinant nucleic acid that comprises at least one subsequence from each of the at least two parental nucleic acids.

The set of fragmented parental nucleic acids recombined in this method are optionally partially produced by cleaving the two parental nucleic acids with a DNase enzyme or by standard synthetic approaches (e.g., automated oligonucleotide synthesis). As another alternative, at least a portion of the set of fragmented parental nucleic acids are optionally produced by partial chain elongation using a polymerase, and one or both of the parental nucleic acids used as templates for elongation of one or more hybridized polymerase primer nucleic acids. Additionally, at least a portion of the set of fragmented parental nucleic acids are optionally produced by synthesizing oligonucleotides which correspond to one or more of the at least two parental nucleic acids, which oligonucleotides include a plurality of codon-

varied oligonucleotides. The at least two parental nucleic acids to be recombined by this method are optionally homologous or non-homologous.

The hybridization step of this method of recombining at least two parental nucleic acids optionally includes hybridizing at least one codon-varied oligonucleotide with at least one additional codon-varied oligonucleotide (e.g., an overlapping codon-varied oligonucleotide) to provide the at least one hybridized nucleic acid. The hybridizing step, alternatively, includes hybridizing at least one codon-varied oligonucleotide with at least one fragmented parental nucleic acid (e.g., DNase fragmented, chemically fragmented, physically fragmented, or the like) to provide the at least one hybridized nucleic acid. As a further option, the hybridizing step can include hybridizing at least one fragmented parental nucleic acid with at least one additional fragmented parental nucleic acid to provide the at least one hybridized nucleic acid.

In a third aspect, the present invention provides a method of recombining homologous or non-homologous nucleic acid sequences having low sequence similarity. The method includes recombining at least one set of fragmented nucleic acids with a set of cross-over codon-varied oligonucleotides, which oligonucleotides individually comprise a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity to produce a recombinant nucleic acid. The resulting recombinant nucleic acid is optionally selected for at least one desired trait or property.

This method of recombining sequences having low sequence similarity optionally includes fragmenting at least one of the homologous or non-homologous nucleic acids to provide the set of fragmented nucleic acids. The homologous or non-homologous nucleic acids are optionally fragmented with a DNase enzyme. The set of fragmented nucleic acids is also optionally provided by synthesizing a plurality of oligonucleotide fragments corresponding to at least one homologous or non-homologous nucleic acid.

A fourth aspect of this invention is a method of recombining a plurality of parental nucleic acids. This method includes ligating a set of a plurality of codon-varied oligonucleotides with a set comprising a plurality of nucleic acid sequences corresponding to a plurality of the parental nucleic acids to produce at least one

recombinant nucleic acid encoding a full-length protein. The set includes at least a first oligonucleotide that is complementary to at least a first of the parental nucleic acids at a first region of sequence diversity and at least a second oligonucleotide which is complementary to at least a second of the parental nucleic acids at a second region of diversity.

Other features of this method include optionally ligating the set of a plurality of oligonucleotides with a ligase. The set of a plurality of oligonucleotides is optionally hybridized to a first parental nucleic acid and ligated with a ligase. Also, the plurality of parental nucleic acids is optionally homologous. Furthermore, the set of a plurality of oligonucleotides optionally comprises a set of codon-varied oligonucleotides (e.g., overlapping codon-varied oligonucleotides). Finally, the method optionally includes hybridizing the set of a plurality of codon-varied oligonucleotides to at least one of the plurality of parental nucleic acids, elongating the oligonucleotides with a polymerase and ligating the resulting elongated oligonucleotides to produce a nucleic acid encoding a substantially full-length protein.

A fifth aspect of the invention relates to various compositions relevant to the methods described herein, such as libraries produced by the methods, recombination mixture compositions, and the like.

A sixth aspect of the present invention is an integrated system that optionally includes a computer or computer readable medium and character strings in a data set that represent a set of codon-varied oligonucleotides (e.g., a set of overlapping codon-varied oligonucleotides). This system optionally integrates a standard automatic synthesizer that is coupled to an output of the computer or computer readable medium. The automatic synthesizer accepts instructions from the computer or computer readable medium and those instructions, in turn, direct the synthesis of a desired set of codon-varied oligonucleotides. Additionally, the automated synthesizer system optionally integrates one or more robotic control elements for, e.g., incubating, denaturing, hybridizing, and elongating the set of oligonucleotides. This version of the integrated system optionally further includes a detector for, e.g., detecting an elongated nucleic acid.

Definitions

Unless otherwise indicated, the following definitions supplement those in the art.

A set of "codon-varied oligonucleotides" is a set of oligonucleotides, similar in sequence but with one or more base variations, where the variations correspond, e.g., to at least one encoded amino acid difference. For example, two codon-varied oligonucleotides can be identical at all codon positions except one which encodes different amino acids. The oligonucleotides are synthesized, e.g., utilizing trinucleotide, i.e., codon-based coupling chemistry. Codon-varied oligonucleotide sequences can be based upon sequences of a selected set of homologous nucleic acids, where the oligonucleotide sequences can include regions of sequence identity and regions of sequence diversity with one or more of those homologous nucleic acids. Aside from being based upon homologous nucleic acid sequences, codon-varied oligonucleotide sequences can also be derived from non-homologous nucleic acids, or a combination of homologous and non-homologous sequences. "Sets" include a plurality of different members, e.g., 2, 3, 4, 5, 10, 20, 50, 100, 1,000 or more different members.

A "consensus region" sequence or subsequence is a region of a polynucleotide having a generalized sequence in which each nucleotide position represents the base most often found in actual sequence comparisons between homologous nucleic acids.

Two nucleic acids "correspond" when they have identical or complementary sequences, when one nucleic acid is a subsequence of the other, or when one sequence is derived naturally or artificially from the other.

A "cross-over" codon-varied oligonucleotide has regions of sequence identity with at least two members of a selected set of nucleic acids that are either homologous or non-homologous.

A "DNase enzyme" is an enzyme that catalyzes the cleavage of DNA, *in vitro* or *in vivo*. Many varieties of DNase enzymes are well characterized, e.g., in Berger and Kimmel, *Guide to Molecular Cloning Techniques, Methods in Enzymology* volume 152 Academic Press, Inc., San Diego, CA; Sambrook *et al.*, *Molecular Cloning - A Laboratory Manual* (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 and *Current Protocols in Molecular Biology*, F.M. Ausubel *et al.*, eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998), and many are commercially available.

Nucleic acids are "elongated" in a reaction that incorporates additional nucleotides, or analogs thereof, or nucleic acids into the nucleic acid sequence. The reaction is typically catalyzed by a polymerase, e.g., a DNA polymerase or a ligase.

A set of "fragmented" nucleic acids results from the cleavage of at least one parental nucleic acid, e.g., enzymatically or chemically, or by providing subsequences of parental sequences in any other manner, including partially elongating a complimentary sequence with a polymerase or utilizing any synthetic format.

A "full-length protein" is a protein with substantially the same sequence domains as a corresponding protein encoded by a natural gene. Such a protein can have altered sequences relative to the corresponding naturally encoded gene, e.g., due to recombination and selection, but unless specified to the contrary, is typically at least about 95% the length of the naturally encoded gene.

Two nucleotide regions have high "sequence similarity" when one region is 90% or more identical to a second selected region when aligned for optimal correspondence. In contrast, regions of low "sequence similarity" refers to those regions that are at most 60% identical, more preferably, 40% or less identical, when aligned for maximal correspondence. Alignment may be accomplished manually or using a common alignment algorithm, such as, e.g., BLAST (set to default parameters).

Nucleic acids are "homologous" when they share sequence similarity that is derived, naturally or artificially, from a common ancestral sequence. This occurs naturally as two or more descendent sequences deviate from a common ancestral sequence over time as the result of mutation and natural selection. Artificially homologous sequences may be generated in various ways. For example, a nucleic acid sequence can be synthesized *de novo* to yield a nucleic acid that differs in sequence from a selected parental nucleic acid sequence. Artificial homology can also be created by artificially recombining one nucleic acid sequence with another, as occurs, e.g., during cloning or chemical mutagenesis, to produce a homologous descendent nucleic acid.

It is generally assumed that the two nucleic acids have common ancestry when they demonstrate sequence similarity. However, the exact level of sequence similarity necessary to establish homology varies in the art. In general, for

purposes of this disclosure, two nucleic acid sequences are deemed to be homologous when they share enough sequence identity to permit direct hybridization-mediated recombination to occur between the two sequences.

It should be noted, however, that a specific advantage of this invention is the capacity to recombine nucleic acids that are more distantly related than some other methods of recombination permit. In this aspect of the invention, nucleic acid sequences that are only distantly related, or not even detectably related, can be recombined by means of cross-over codon-varied oligonucleotides which are described, *supra*.

Nucleic acids "hybridize" when complementary single strands of nucleic acid pair to give a double-stranded nucleic acid sequence. Hybridization occurs due to a variety of well-characterized forces, including hydrogen bonding, solvent exclusion, and base stacking. An extensive guide to nucleic hybridization may be found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology—Hybridization with Nucleic Acid Probes*, part I, chapter 2, "Overview of principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, New York."

A "library" is a set of oligonucleotides. The set can be pooled, or can be individually accessible. The oligonucleotides may comprise DNA, RNA or combinations thereof.

Nucleic acid sequences are "overlapping" when they possess at least one complementary subsequence.

Nucleic acids are "non-homologous" when they lack shared sequence similarity with a common ancestral sequence, or when they can only be indirectly recombined utilizing oligonucleotide intermediates.

A nucleic acid "domain" is a discrete nucleic acid region or subsequence. It may be conserved or not conserved between a plurality of homologous nucleic acids. Generally, a domain is specified by comparing two or more sequences, where regions of sequence diversity between sequences constitutes a "sequence diversity domain," while a region of similarity is a "sequence similarity domain."

Two nucleic acids "recombine" when sequences or subsequences from each of the two nucleic acids are combined in a progeny nucleic acid. Two sequences

are "directly" recombined when both are substrates for recombination. Two sequences are "indirectly" recombined when the sequences are recombined by means of an intermediate such as a cross-over codon-varied oligonucleotide. When two nucleic acid sequences indirectly recombine, no more than one of those sequences is an actual substrate for recombination, and in some cases, neither sequence is a substrate for recombination.

A "solid-phase substrate sequence" typically comprises at least one nucleotide covalently attached at its 3' end to a solid support, or another chemical group to which nucleotides can be attached.

The term "trinucleotide phosphoramidite sequence" is any codon sequence of nucleotides synthesized using standard phosporamidite chemistry. Many sources have described such synthesis, e.g., Virnekäs, B. *et al.*, (1994) *Nucleic Acids Res.*, 22, 5600-5607 and Kayushin, A. L. *et al.*, (1996) *Nucleic Acids Res.*, 24, 3748-3755.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates a possible hidden Markov model for the peptide ACCY.

DETAILED DISCUSSION OF THE INVENTION

INTRODUCTION

The present invention provides recombination methodologies in which codon-varied oligonucleotides are used to provide recombined nucleic acid populations. Codon-varied oligonucleotides are chemically synthesized, e.g., utilizing trinucleotide phosphoramidite sequences and/or mononucleotides. These oligonucleotides can be derived from homologous or non-homologous nucleic acid sequences, or combinations of such sequences. In general, the use of libraries of codon-varied oligonucleotides for recombination and/or gene synthesis significantly enhances the rate of recombination processes. Furthermore, codon-varied oligonucleotide intermediates may be used to achieve indirect recombination of nucleic acid sequences that would not otherwise recombine.

Mononucleotide-based oligonucleotide synthesis has significant limitations relative to trinucleotide-based formats. One major limitation presented by mononucleotide formats, when adding diversity, stems from the degeneracy of the

genetic code. For example, if serine (via UCC only) or glycine (via GGC only) is desired at a specific position, mononucleotide-based synthesis leads to a degenerate oligonucleotide encoding the amino acid sequence, lysine-serine-serine (i.e., KSS). However, the KSS-encoding oligonucleotide also leads to the insertion of undesired amino acids, e.g., as follows: tryptophan (UGG), cysteine (UGC), and alanine (GCC). In contrast, a trinucleotide-generated codon-varied oligonucleotide results solely in the preferred insertions, because nucleotides are added to the sequence being synthesized as part of pre-defined codon-length units. Thus, trinucleotide-based techniques afford greater control over polynucleotide synthetic processes than monomer-based approaches.

Prior to the present invention, trinucleotide phosphoramidites had not been used to synthesize codon-varied oligonucleotides for use in DNA recombination. Advantages of this method include being able to recombine DNA at various levels, e.g., as defined polynucleotide fragments or as individual codon sequences. Amino acids at any position can be designated specifically, incorporated in a biased manner, or inserted randomly. Additionally, being codon-based, any deletions and insertions, whether intended or due to error, that may occur during synthesis, will not offset the coding frame of reference and in turn, will be less likely to inactivate encoded recombined proteins. As such, the present invention will enhance the capacity of recombination-based artificial evolution techniques.

In overview, the present invention initially entails determining the specific nucleic acid sequences that are to be synthesized as codon-varied oligonucleotides and in turn, which form compositions for recombination. Nucleic acid synthesis, as well as other aspects of the invention, can be conducted in a fully integrated system that incorporates a computer coupled to an automatic synthesizer and one or more robotic control elements. Oligonucleotides can be synthesized using either a trinucleotide coupling format or by mononucleotide synthesis in a split-pool format. Following synthesis, recombination can be carried out using one of several alternative methods. Finally, desired traits or properties can be selected using techniques that are known in the art. These recombination-based evolution methods are optionally combined with any available recombination/mutation approach to further increase diversity of resulting nucleic acids. For example, these alternative nucleic acid diversification methods can be applied before or after selection.

Furthermore, any of these procedures can be repeated reiteratively whether separately or in combination.

The following provides details regarding the various aspects of the recombination methods of the present invention, including synthesis, hybridization, elongation, and selection protocols. It also provides details regarding the different compositions and integrated systems of the present invention.

SELECTION OF HOMOLOGOUS/NON-HOMOLOGOUS NUCEIC ACID SEQUENCES TO BE SYNTHESIZED AS CODON-VARIED OLIGONUCLEOTIDES FOR RECOMBINATION

A threshold issue in practicing the present invention is selecting or designing the sequences of the codon-varied oligonucleotides to be synthesized. They can be derived from nucleic acid sequences that are homologous, non-homologous, and/or purely practitioner designed. In an aspect of the invention, designated mixtures of trinucleotide phosphoramidites can be used to vary an amino acid at any position according to any desired specifications. Additionally, positions may be made either random or biased by any pattern, and for family shuffling each amino acid position can be varied with respect to any known natural or artificial diversity in the sequences under consideration. Also, deletions and insertions can be programmed or the reaction conditions can be adjusted so that they occur at some frequency. In another aspect, multiple natural and/or designed parent sequences having defined motifs can be synthesized. According to this method, it is useful to create hybrid parents that together contain the sequence elements of the parents, but which are "pre-shuffled" so as to promote recombination with the natural parents.

Sequence information available from nucleic acid databases is useful references during the selection and design process. Genbank®, Entrez®, EMBL, DDBJ, and the NCBI are examples of public database/search services that can be accessed. Many sequence databases are available via the internet or on a contract basis from a variety of companies specializing in genomic information generation and/or storage.

When recombining homologous nucleic acids, the present invention optionally includes aligning homologous nucleic acid sequences or regions of similarity. For example, in one aspect, the invention relates to a method of recombining at least two parental nucleic acids. In an embodiment of this method, the

composition of nucleic acids to be recombined is provided by aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. Codon-varied oligonucleotides are then synthesized to correspond to at least one region of sequence diversity. Similarly, an aspect of the invention includes deriving the sequences of an additional set of nucleic acid fragments from first round selected nucleic acids by aligning those first round sequences to identify regions of identity and regions of diversity.

In these processes of sequence comparison and homology determination, one sequence is often used as a reference against which other test nucleic acid sequences are compared. This comparison can be accomplished with the aid of a sequence comparison algorithm, i.e., instruction set, or by visual inspection. When an algorithm is employed, test and reference sequences are input into a computer, subsequence coordinates are designated, as necessary, and sequence algorithm program parameters are specified. The algorithm then calculates the percent sequence identity for the test nucleic acid sequence(s) relative to the reference sequence, based on the specified program parameters.

For purposes of the present invention, suitable sequence comparisons can be executed, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see generally*, Ausubel *et al., supra*).

One example algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al., J. Mol. Biol.* 215:403-410 (1990). *See also*, Altschul *et al., Nucleic Acids Res.* 25(17):3389. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the

neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul (1993) *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence (and, therefore, in this context, homologous) if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, or less than about 0.01, and or even less than about 0.001.

An additional example of a useful sequence alignment algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. It can also plot a tree showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng & Doolittle, *J. Mol. Evol.* 35:351-360

(1987). The method used is similar to the method described by Higgins & Sharp, *CABIOS*5:151-153 (1989). The program can align, e.g., up to 300 sequences of a maximum length of 5,000 letters. The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster can then be aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences can be aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments. The program can also be used to plot a dendogram or tree representation of clustering relationships. The program is run by designating specific sequences and their amino acid or nucleotide coordinates for regions of sequence comparison.

An example of aligning proteins of unknown function relies on the flexible statistical model called the hidden Markov model (HMM). This model utilizes defined 'threading' through a multiple sequence alignment. The threading matrix, but not the sequence alignment consensus itself, is subsequently used to identify novel proteins that can be clustered with the original group of HMM structures. HMM or variations thereof can be excellent statistical tools used in defining sequences and divergence in split-pool synthesis of codon-varied oligonucleotides. Using the HMM matrix, each position is given a certain set of options (e.g., delete, insert or add the next amino acid) and the percentage of codon-varied oligonucleotide-carrying beads going down each path can easily be calculated based on a parental display.

As shown in Figure 1, the typical profile hidden Markov model is a chain of match (square), insert (diamond), and delete (circle) nodes, with all transitions between nodes and all character costs in the insert and match nodes trained to specific probabilities (i.e., the known parents). The single best path through an HMM corresponds to a path from a start state to an end state in which each character of the sequence is related to a successive match or insertion state along that path. Delete states indicate that the sequence has no character corresponding to that position in the HMM.

Transitions from state to state progress from left to right through the model, with the exception of self-loops on insertion states (FIG. 1). The self-loops

allow deletions of any length to fit the model, regardless of the length of other sequences in the family.

A path through the model can represent any sequence. The probability of any sequence, given the model, is computed by multiplying the emission and transition probabilities along the path. Figure 1 illustrates a possible hidden Markov model for the peptide ACCY. As shown, a path through the model represented by ACCY is highlighted. The peptide is represented as a sequence of probabilities. The numbers in the boxes show the probability that an amino acid occurs in a particular state, and the number next to the directed arcs show probabilities which connect the states. For instance, the probability of A being emitted in position one is 0.3, and the probability of C being emitted in position two is 0.6. The probability of ACCY along this path is:

$$.4 * .3 * .46 * .6 * .97 * .5 * .015 * .73 * .01 * 1 = 1.76 \times 10^{-6}.$$

Or, by transforming probabilities to logs so that addition can replace multiplication:

$\log_e(.4) + \log_e(.3) + \log_e(.46) + \log_e(.6) + \log_e(.97) + \log_e(.5) + \log_e(.015) + \log_e(.73) + \log_e(.01) + \log_e(1) = -13.25.$

CODON-VARIED OLIGONUCLEOTIDE SYNTHESIS

One aspect of the present invention comprises synthesizing codon-varied oligonucleotides which are then used to achieve recombination. Codon-varied oligonucleotides can be synthesized, e.g., utilizing trinucleotide-based phosphoramidite coupling chemistry, in which trinucleotide phosphoramidites representing codons for all 20 amino acids are used to introduce entire codons into oligonucleotide sequences synthesized by this solid-phase technique. An advantage of this trinucleotide synthetic approach is that it provides tremendous flexibility to recombination processes as codon-varied oligonucleotide sequences can be selected or designed by the practitioner, e.g., to be based upon homologous or non-homologous nucleotide sequences, or combinations of such sequences. Optionally, oligonucleotides of the present invention are synthesized using mononucleotide approaches known in the art, either alone or in conjunction with trinucleotide-based synthetic techniques.

The synthesis of trinucleotide phoshoramidites, their subsequent use in oligonucleotide synthesis, and related issues are described in, e.g., Virnekäs, B., *et al.*,

(1994) *Nucleic Acids Res.*, 22, 5600-5607, Kayushin, A. L. *et al.*, (1996) *Nucleic Acids Res.*, 24, 3748-3755, Huse, U.S. Pat. No. 5,264,563 "PROCESS FOR SYNTHESIZING OLIGONUCLEOTIDES WITH RANDOM CODONS," Lyttle *et al.*, U.S. Pat. No. 5,717,085 "PROCESS FOR PREPARING CODON AMIDITES," Shortle *et al.*, U.S. Pat. No. 5,869,644 "SYNTHESIS OF DIVERSE AND USEFUL COLLECTIONS OF OLIGONUCLEOTIDES," Greyson, U.S. Pat. No. 5,789,577 "METHOD FOR THE CONTROLLED SYNTHESIS OF POLYNUCLEOTIDE MIXTURES WHICH ENCODE DESIRED MIXTURES OF PEPTIDES," and Huse, WO 92/06176 "SURFACE EXPRESSION LIBRARIES OF RANDOMIZED PEPTIDES."

In the present invention, codon-varied oligonucleotides can be synthesized using various trinucleotide-related techniques, e.g., the trinucleotide synthesis format and/or the split-pool synthesis format. The individual steps for performing both formats are described, *infra*. Preferably, all of the oligonucleotides of a selected length (e.g., about 20, about 30, about 40, about 50, about 60, about 70, about 80, about 90, or about 100 or more nucleotides) which incorporate the chosen nucleic acid sequences are synthesized.

In general, overlapping codon-varied oligonucleotides, synthesized according to the methods of the present invention, can have, e.g., about 10 bases of sequence identity to either side of a region of variance to ensure reasonably efficient recombination. However, flanking regions with identical bases can have fewer identical bases (e.g., about 5, about 6, about 7, about 8, or about 9) and can, of course, have larger regions of identity (e.g., about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19, about 20, about 25, about 30, about 50, or more bases).

The trinucleotide synthesis format includes providing a solid-phase substrate sequence having a 5' terminus and at least one base, both of which have protecting groups thereon. The 5' protecting group of the solid-phase substrate sequence is then removed to provide a 5' deprotected solid-phase substrate sequence, which is then coupled with a selected trinucleotide phosphoramidite sequence. The trinucleotide has a 3' terminus, a 5' terminus, and three bases, each of which has protecting groups thereon. The coupling step yields an extended oligonucleotide sequence. Thereafter, the removing and coupling steps are optionally repeated.

When these steps are repeated, the extended oligonucleotide sequence yielded by each repeated coupling step becomes the solid-phase substrate sequence of the next repeated removing step until a desired codon-varied oligonucleotide is obtained. This basic synthesis format can optionally include coupling together one or more of: mononucleotides, trinucleotide phosphoramidite sequences, and oligonucleotides.

The split-pool synthesis format includes providing solid-phase substrate sequences, each having a 5' terminus and at least one base, both of which have protecting groups thereon. The 5' protecting groups of the solid-phase substrate sequences are removed to provide 5' deprotected solid-phase substrate sequences, which are then coupled with selected trinucleotide phosphoramidite sequences. Each trinucleotide has a 3' terminus, a 5' terminus, and three bases, all of which have protecting groups thereon. The coupling step yields extended oligonucleotide sequences. Thereafter, the removing and coupling steps are optionally repeated. When these steps are repeated, the extended oligonucleotide sequences yielded by each repeated coupling step become the solid-phase substrate sequences of the next repeated removing step until extended intermediate oligonucleotide sequences are produced.

Additional steps of the split-pool format optionally include splitting the extended intermediate oligonucleotide sequences into two or more separate pools. After this is done, the 5' protecting groups of the extended intermediate oligonucleotide sequences are removed to provide 5' deprotected extended intermediate oligonucleotide sequences in the two or more separate pools. Following this, these 5' deprotected intermediates are coupled with one or more selected mononucleotides, trinucleotide phosphoramidite sequences, or oligonucleotides in the two or more separate pools to yield further extended intermediate oligonucleotide sequences. In turn, these further extended sequences are pooled into a single pool. Thereafter, the steps beginning with the removal of the 5' protecting groups of the solid-phase substrate sequences to provide 5' deprotected solid-phase substrate sequences are optionally repeated. When these steps are repeated, the further extended oligonucleotide sequences, yielded by each repeated coupling step that generates those specific sequences, become the solid-phase substrate sequences of the next repeated removing step that includes those specific sequences until desired codon-varied oligonucleotides are obtained.

The split-pool synthesis format is particularly advantageous when sequences having low homology (e.g., <90%) are to be synthesized. For example, it enables the practitioner to precisely control the way sequences being elongated are dispersed within a framework of defined nucleic acid sequences. Utilizing this format, one can now take co-variance among divergent parental nucleic acids into account by being able to define the exact point where recombination is to occur and what percentage of those parents will form a chimera at that point. Further, co-evolution and systematic variation within a nucleic acid sequence can be captured using this module-based format. In this approach, the smallest module within each genetic sequence is the codon, i.e., a single trinucleotide in length, while larger modules are, e.g., at least 15-90 or more nucleotides in length, which can be a structurally defined segment, or a sequence of less or no homology to related parent nucleic acids. Also, one can foresee the use of a random or Poisson distribution of sequence sizes to initiate the recombination process. This would mimic current recombination methods, but in a synthetic and controlled format.

The chemistry involved in both the trinucleotide and the split-pool codon-varied oligonucleotide synthetic method is known to those of skill. In general, both methods optionally utilize phosphoramidite solid-phase chemical synthesis in which the 3' ends of nucleic acid solid-phase substrate sequences are optionally covalently attached to a solid support, e.g., controlled pore glass. The 5' protecting groups can be, e.g., a triphenylmethyl group, such as, dimethoxyltrityl (DMT) or monomethyoxytrityl, a carbonyl-containing group, such as, 9-fluorenylmethyloxycarbonyl (FMOC) or levulinoyl, an acid-clearable group, such as, pixyl, a fluoride-cleavable alkylsilyl group, such as, tert-butyl dimethylsilyl (T-BDMSi), triisopropyl silyl, or trimethylsilyl. The 3' protecting groups can be, e.g., β-cyanoethyl groups.

Both synthesis formats can optionally be performed in an integrated automated synthesizer system that automatically performs the synthetic steps. This aspect includes inputting character string information into a computer, the output of which then directs the automated synthesizer to perform the steps necessary to synthesize the desired codon-varied oligonucleotides. This integrated system is discussed further, *infra*.

To further ensure that functional, full-length recombined genes are ultimately obtained, certain techniques can be utilized following codon-varied oligonucleotide synthesis. For example, gel purification is one method that can be used to purify synthesized oligonucleotides. High-performance liquid chromatography can be similarly employed.

Following synthesis, translational coupling can be used to assess gene functionality, e.g., to test whether full-length sequences are generated. In this process, the translation of a reporter protein, e.g., green fluorescent protein or β-galactosidase is coupled to that of the recombined gene product. This enables one to distinguish full-length recombined genes from those that contain deletions or frame shifts. The subsequent selection of desired traits or properties of the recombined gene is discussed further, *infra*.

The various references already discussed which relate to oligonucleotide synthesis provide further details on synthesis of oligonucleotides by either trinucleotide or mononucleotide chemical synthesis.

DIVERSITY GENERATING METHODS OF THE PRESENT INVENTION

The present invention provides several methods for recombining nucleic acid sequences. In one aspect, the invention is directed to a method of recombining codon-varied oligonucleotides to provide a population of recombined nucleic acids. Following oligonucleotide synthesis, this method comprises hybridizing and elongating a set of codon-varied oligonucleotides (e.g., overlapping codon-varied oligonucleotides) to provide the population of recombined nucleic acids. The invention also provides a method of recombining at least two parental nucleic acids to provide at least one recombinant nucleic acid. Beyond providing a composition comprising at least one set of fragmented parental nucleic acids, the composition is similarly hybridized and the hybridization product is then elongated to provide recombinant nucleic acids that comprise at least one subsequence from each of the two parental nucleic acids. Another method of the invention is that of recombining homologous or non-homologous nucleic acid sequences having low sequence similarity. This method comprises, e.g., recombining at least one set of fragmented nucleic acids with a set of cross-over codon-varied oligonucleotides. This recombination method also involves hybridization and elongation steps.

In Vitro Recombination

According to certain methods of the invention, codon-varied oligonucleotides can be recombined *in vitro*, e.g., in a pool of such sequences. For example, a set of single-stranded codon-varied oligonucleotides can be synthesized, with individual members having sequences that are complementary to one another. Such single-stranded sequences can then be hybridized, e.g., by cooling to about 20°C to about 75°C, and preferably from about 40°C to about 65°C. Hybridization can be accelerated by the addition of polyethylene glycol ("PEG") or salt. The salt concentration is, e.g., from about 0 mM to about 600 mM, or, e.g., from about 10 mM to about 100 mM. The salt can be such salts as $(NH_4)_2SO_4$, KCl, NaCl, or the like. The concentration of PEG is preferably from about 0% to about 20%, more preferably from about 5% to about 10%.

During elongation, the hybridized codon-varied oligonucleotides are then incubated in the presence of a nucleic acid polymerase, e.g., Taq or Klenow, and dNTP's (i.e., dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq or other high-temperature polymerase can be used with a hybridization temperature of between about 45° C to about 65°C. If the areas of identity are small, Klenow or other low-temperature polymerases can be used with a hybridization temperature of between about 20°C to about 30°C. The polymerase can be added to the random nucleic acid fragments prior to, simultaneously with, or after hybridization. As noted elsewhere in this disclosure, certain embodiments of the invention can involve denaturing the resulting elongated double-stranded nucleic acid sequences and then hybridizing and elongating those sequences again. This cycle can be repeated for any desired number of times. Preferably the cycle is repeated from about 2 to about 100 times, e.g., from about 10 to about 40 times.

In Vivo Recombination

An embodiment of the present invention involves *in vivo* recombination. In this embodiment, a population of codon-varied oligonucleotides can be introduced into bacterial or eukaryotic cells under conditions such that at least one codon-varied oligonucleotide sequence is present in each host cell. Oligonucleotide sequences can be introduced into host cells using various methods known in the art, e.g., calcium chloride treatment, electroporation, transfection, lipofection, biolistics, conjugation, and the like. *In vivo* recombination formats that can optionally be used in the present invention are, e.g., plasmid-plasmid

recombination, virus-plasmid recombination, virus-virus recombination, chromosome recombination, virus-chromosome recombination, chimeric recombination in which the codon-varied oligonucleotides are chimeraplasts, and the like. For example, when two oligonucleotide sequences that have regions of identity are inserted into the host cells homologous recombination occurs between the two sequences.

After transformation, the host cell transformants are placed under selection to identify those host cell transformants which contain specific nucleic acid sequences having desired traits or properties. For example, if increased resistance to a particular drug is desired then the transformed host cells may be subjected to increased concentrations of the particular drug and those transformants producing mutated proteins able to confer increased drug resistance will be selected. If the enhanced ability of a particular protein to bind to a receptor is desired, then expression of the protein can be induced from the transformants and the resulting protein assayed in a ligand binding assay by methods known in the art to identify that subset of the recombined population which shows enhanced binding to the ligand. Alternatively, the protein can be expressed in another system to ensure proper processing. The steps of this process can be repeated for multiple cycles.

The host cells can also be recursively recombined via, e.g., protoplast fusion and other whole genome recombination methodologies to increase the diversity of the cell populations. Whole genome recombination methods are described in detail in International Application Nos. PCT/US98/00852 and PCT/US99/15972, filed January 16, 1998 and July 15, 1999, respectively

Recombination by Ligation

One aspect of the present invention relates to a method of performing recombination between nucleic acids by ligation of libraries of codon-varied oligonucleotides corresponding to the nucleic acids to be recombined. In this format, a set of a plurality of codon-varied oligonucleotides which includes a plurality of nucleic acid sequences from a plurality of the parental nucleic acids are ligated to produce a recombinant nucleic acid, typically encoding a full length protein (although ligation can also be used to make libraries of partial nucleic acid sequences which can then be recombined, e.g., to produce a partial or full-length recombinant nucleic acid via ligation or polymerase-mediated methods). The oligonucleotide set typically includes at least a first oligonucleotide which is complementary to at least a first of

the parental nucleic acids at a first region of sequence diversity and at least a second oligonucleotide which is complementary to at least a second of the parental nucleic acids at a second region of diversity. The parental nucleic acids can be homologous or non-homologous.

Typically, the codon-varied oligonucleotides are ligated with a ligase. In one general format, the oligonucleotides are hybridized to a first parental nucleic acid which acts as a template, and ligated with a ligase. The codon-varied oligonucleotides may be extended with a polymerase and ligated. The polymerase can be, e.g., an ordinary DNA polymerase or a thermostable DNA polymerase. The ligase can also be, e.g., an ordinary DNA ligase or a thermostable DNA ligase. Many such polymerases and ligases are commercially available.

In one set of approaches to recombination, a common element is the preparation of a single-stranded (ss) template to which codon-varied oligonucleotide primers are annealed and then elongated by a DNA polymerase in the presence of dNTP's and an appropriate buffer. The gapped duplex can be sealed with ligase prior to, e.g., transformation or electroporation into *E. coli.*, where the newly synthesized strand is replicated and generates a chimeric gene with contributions from the codon-varied oligonucleotide in the context of the single-stranded (ss) parent.

The ss template to which codon-varied oligonucleotides can be annealed can be prepared, for example, by the incorporation of the phage IG region into a plasmid and use of a helper phage such as M13KO7 (Pharmacia Biotech) or R408 to package ss plasmids into filamentous phage particles. Optionally, the ss template can be generated by denaturation of a double-stranded (ds) template and annealing in the presence of the codon-varied oligonucleotide primers. A variety of additional ligation-based approaches are described in, e.g., "SINGLE-STRANDED NUCLEIC ACID TEMPLATE-MEDIATED RECOMBINATION AND NUCLEIC ACID FRAGMENT ISOLATION" by Affholter, filed Sept. 6, 2000 (USSN 09/656,549) and, e.g., "Methods and Compositions for Polypeptide Engineering" by Stemmer et al. (WO 98/27230).

Enrichment methods vary for isolating newly synthesized chimeric strand over the parental template strand. Isolation and selection of ds templates can be performed using available methods. See e.g., Ling et al. (1997) "Approaches to DNA mutagenesis: an overview" *Anal Biochem.*, Dec 15;254(2):157-78; Dale et al.

(1996) "Oligonucleotide-directed random mutagenesis using the phosphorothioate method" *Methods Mol Biol.*, 57:369-74; and Smith (1985) "In vitro mutagenesis" *Ann. Rev. Genet.*, 19:423-462.

In one aspect, for example, a "Kunkel style" method uses uracil containing templates. Similarly, the "Eckstein" method uses phosphorothioate-modified DNA (Taylor et al. (1985) "The use of phosphorothioate-modified DNA in restriction enzyme reactions to prepare nicked DNA." *Nucleic Acids Res.* 13:8749-8764; Taylor et al. (1985) "The rapid generation of oligonuclcotide-directed mutations at high frequency using phosphorothioate-modified DNA" *Nucleic Acids Res.* 13:8765-8787; Nakamaye & Eckstein (1986) "Inhibition of restriction endonuclease Nci I cleavage by phosphorothioate groups and its application to oligonucleotide-directed mutagenesis." *Nucleic Acids Res.* 14: 9679-9698; Sayers et al. (1988). "Y-T Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis." *Nucleic Acids Res.* 16:791-802; Sayers et al. (1988) "5'-3' Strand specific cleavage of phosphorothioate-containing DNA by reaction with restriction endonucleases in the presence of ethidium bromide" *Nucleic Acids Res.* 16:803-814). The use of restriction selection, or e.g., purification can be used in conjunction with mismatch repair deficient strains (*see, e.g.*, Carter et al. (1985) "Improved oligonucleotide site directed mutagenesis using M13 vectors" *Nucleic Acids Res.* 13, 4431-4443 Carter (1987) "Improved oligonucleotide-directed mutagenesis using M13 vectors." *Methods in Enzymol.* 154:382-403; Wells (1986) "Importance of hydrogen bond formation in stabilizing the transition state of subtilisin." *Trans. R. Soc. Lond.* A317, 415-423).

The "mutagenic" primers used in these methods can be codon-varied oligonucleotide(s) encoding, e.g., any type of randomization, insertion, deletion based on sequence diversity of homologous genes, etc. Multiple codon-varied oligonucleotide primers can anneal to a given template and be extended to create multiply chimeric genes. The use of a DNA polymerase such as those from phages T4 or T7 are suitable for this purpose as they do not degrade or displace a downstream primer from the template.

In one example, DNA recombination is performed using uracil containing templates. In this embodiment, the gene of interest is cloned into an *E. coli* plasmid containing the filamentous phage intergenic (IG, ori) region. Single

stranded (ss) plasmid DNA is packaged into phage particles upon infection with a helper phage such as M13KO7 (Pharmacia) or R408 and can be easily purified by methods such as phenol/chloroform extraction and ethanol precipitation. If this DNA is prepared in a dut- ung- strain of *E. coli*, a small number of uracil residues are incorporated into it in place of the normal thymine residues. One or more codon-varied oligonucleotide primers are annealed to the ss uracil-containing template by heating to 90°C and slowly cooling to room temperature. An appropriate buffer containing all 4 dNTPs, T7 DNA polymerase and T4 DNA ligase is added to the annealed template/primer mix and incubated between room temperature-37°C for ≥1 hour. The T7 DNA polymerase extends from the 3' end of the codon-varied oligonucleotide primer and synthesizes a complementary strand to the template incorporating the primer. DNA ligase seals the gap between the 3' end of the newly synthesized strand and the 5' end of the primer. If multiple codon-varied oligonucleotide primers are used, then the polymerase will extend to the next primer, stop and ligase will seal the gap. This reaction is then transformed into an ung+ strain of *E. coli* and antibiotic selection for the plasmid is applied. The uracil N-glycosylase (ung gene product) enzyme in the host cell will recognize the uracil in the template strand and removes it, creating apyrimidinic sites that are either not replicated or the host repair systems will correct it by using the newly synthesized strand as a template. The resulting plasmids predominantly contain the desired change in the gene if interest. If multiple codon-varied oligonucleotide primers are used then it is possible to simultaneously introduce numerous changes in a single reaction. If the codon-varied oligonucleotide primers are derived from or correspond to fragments of homologous genes, then multiply chimeric genes can be generated.

Iterative Codon-Varied Oligonucleotide-Mediated Recombination Methods

In one embodiment, the present invention provides iterative codon-varied oligonucleotide-mediated recombination formats. These formats can be combined with standard recombination- or mutation-based methods, also, optionally, in an iterative format.

In particular, recombinant nucleic acids produced by codon-varied oligonucleotide-mediated recombination can be screened for activity and sequenced. The sequenced recombinant nucleic acids are aligned and regions of identity and

diversity are identified. Codon-varied oligonucleotides are then selected for recombination of the sequenced recombinant nucleic acids. This process of screening, sequencing active recombinant nucleic acids and recombining the active recombinant nucleic acids can be iteratively repeated until a molecule with a desired trait or property is obtained.

In addition, recombinant nucleic acids made using codon-varied oligonucleotides can be cleaved and recombined using available recombination methods, which are, optionally, reiterative. Standard recombination can be used in conjunction with oligonucleotide recombination and either or both steps are optionally reiteratively repeated.

An example of iterative recombination by oligonucleotide-mediated recombination of codon-varied oligonucleotides occurs when extremely fine grain recombination is desired. For example, genes encoding small proteins such as defensins (antifungal proteins of about 50 amino acids), EF40 (an antifungal protein family of about 28 amino acids), peptide antibiotics, peptide insecticidal proteins, peptide hormones, many cytokines and many other small proteins, are difficult to recombine by standard recombination methods, because the recombination occurs with a frequency that is roughly the same as the size of the gene to be recombined, limiting the diversity resulting from recombination. In contrast, oligonucleotide-mediated recombination methods can recombine essentially any region of diversity in any set of sequences, with crossovers occurring at any selected base-pair.

Thus, libraries of sequences prepared by recursive codon-varied oligonucleotide mediated recombination are optionally screened and selected for a desired property, and improved (or otherwise desirable) clones are sequenced with the process being iteratively repeated to generate additional libraries of nucleic acids. Thus, additional recombination rounds are performed either by standard fragmentation-based recombination methods, or by sequencing positive clones, designing appropriate family shuffling oligonucleotides and performing a second round of recombination/selection to produce an additional library (which can be recombined as described). In addition, libraries made from different recombination rounds can also be recombined, either by sequencing and oligonucleotide recombination or by standard recombination methods.

Additional Recombination and Mutagenesis Approaches

A variety of additional diversity generating protocols is available and described in the art. The procedures can be used separately, and/or in combination with any method described herein to produce one or more variants of a nucleic acid or set of nucleic acids, as well as variants of encoded proteins. Individually and collectively, these procedures provide robust, widely applicable ways of generating diversified nucleic acids and sets of nucleic acids (including, e.g., nucleic acid libraries) useful, e.g., for the engineering or rapid evolution of nucleic acids, proteins, pathways, cells and/or organisms with new and/or improved characteristics. These methods can be used in combination with any of the methods described herein (e.g., codon-varied oligonucleotide-mediated recombination, etc.), either to provide substrates for the methods herein, or to further modify, mutate or evolve any recombined nucleic acid produced herein, or both.

While distinctions and classifications are made in the course of the ensuing discussion for clarity, it will be appreciated that the techniques are often not mutually exclusive. Indeed, the various methods can be used singly or in combination, in parallel or in series, with each other or with the methods herein, to generate diverse sequence variants and to screen for desirable activity in such diverse variants.

The result of any of the diversity generating procedures described herein can be the generation of one or more nucleic acids, which can be selected or screened for nucleic acids that encode proteins with or which confer desirable properties. Following diversification by one or more of the methods herein, or otherwise available to one of skill, any nucleic acids that are produced can be selected for a desired activity or property. This can include identifying any activity that can be detected, for example, in an automated or automatable format, by any of the assays in the art as discussed below. A variety of related (or even unrelated) properties can be evaluated, in serial or in parallel, at the discretion of the practitioner.

Descriptions of a variety of diversity generating procedures for modifying nucleic acid sequences are found the following publications and the references cited therein: Stemmer, et al. (1999) "Molecular breeding of viruses for targeting and other clinical properties" *Tumor Targeting* 4:1-4; Ness et al. (1999) "DNA Shuffling of subgenomic sequences of subtilisin" *Nature Biotechnology*

17:893-896; Chang et al. (1999) "Evolution of a cytokine using DNA family shuffling" *Nature Biotechnology* 17:793-797; Minshull and Stemmer (1999) "Protein evolution by molecular breeding" *Current Opinion in Chemical Biology* 3:284-290; Christians et al. (1999) "Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling" *Nature Biotechnology* 17:259-264; Crameri et al. (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution" *Nature* 391:288-291; Crameri et al. (1997) "Molecular evolution of an arsenate detoxification pathway by DNA shuffling," *Nature Biotechnology* 15:436-438; Zhang et al. (1997) "Directed evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening" *Proc. Natl. Acad. Sci. USA* 94:4504-4509; Patten et al. (1997) "Applications of DNA Shuffling to Pharmaceuticals and Vaccines" *Current Opinion in Biotechnology* 8:724-733; Crameri et al. (1996) "Construction and evolution of antibody-phage libraries by DNA shuffling" *Nature Medicine* 2:100-103; Crameri et al. (1996) "Improved green fluorescent protein by molecular evolution using DNA shuffling" *Nature Biotechnology* 14:315-319; Gates et al. (1996) "Affinity selective isolation of ligands from peptide libraries through display on a lac repressor 'headpiece dimer'" *Journal of Molecular Biology* 255:373-386; Stemmer (1996) "Sexual PCR and Assembly PCR" In: The Encyclopedia of Molecular Biology. VCH Publishers, New York. pp.447-457; Crameri and Stemmer (1995) "Combinatorial multiple cassette mutagenesis creates all the permutations of mutant and wildtype cassettes" *BioTechniques* 18:194-195; Stemmer et al., (1995) "Single-step assembly of a gene and entire plasmid form large numbers of oligodeoxy-ribonucleotides" *Gene*, 164:49-53; Stemmer (1995) "The Evolution of Molecular Computation" *Science* 270: 1510; Stemmer (1995) "Searching Sequence Space" *Bio/Technology* 13:549-553; Stemmer (1994) "Rapid evolution of a protein in vitro by DNA shuffling" *Nature* 370:389-391; and Stemmer (1994) "DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution." *Proc. Natl. Acad. Sci. USA* 91:10747-10751.

Mutational methods of generating diversity, which can be practiced in combination with other diversity generation methods including those noted herein, include, for example, site-directed mutagenesis (Ling et al. (1997) "Approaches to DNA mutagenesis: an overview" *Anal Biochem.* 254(2): 157-178; Dale et al. (1996) "Oligonucleotide-directed random mutagenesis using the phosphorothioate method"

Methods Mol. Biol. 57:369-374; Smith (1985) "In vitro mutagenesis" *Ann. Rev. Genet.* 19:423-462; Botstein & Shortle (1985) "Strategies and applications of in vitro mutagenesis" *Science* 229:1193-1201; Carter (1986) "Site-directed mutagenesis" *Biochem. J.* 237:1-7; and Kunkel (1987) "The efficiency of oligonucleotide directed mutagenesis" in *Nucleic Acids & Molecular Biology* (Eckstein, F. and Lilley, D.M.J. eds., Springer Verlag, Berlin)); mutagenesis using uracil containing templates (Kunkel (1985) "Rapid and efficient site-specific mutagenesis without phenotypic selection" *Proc. Natl. Acad. Sci. USA* 82:488-492; Kunkel et al. (1987) "Rapid and efficient site-specific mutagenesis without phenotypic selection" *Methods in Enzymol.* 154, 367-382; and Bass et al. (1988) "Mutant Trp repressors with new DNA-binding specificities" *Science* 242:240-245); oligonucleotide-directed mutagenesis (*Methods in Enzymol.* 100: 468-500 (1983); *Methods in Enzymol.* 154: 329-350 (1987); Zoller & Smith (1982) "Oligonucleotide-directed mutagenesis using M13-derived vectors: an efficient and general procedure for the production of point mutations in any DNA fragment" *Nucleic Acids Res.* 10:6487-6500; Zoller & Smith (1983) "Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13 vectors" *Methods in Enzymol.* 100:468-500; and Zoller & Smith (1987) "Oligonucleotide-directed mutagenesis: a simple method using two oligonucleotide primers and a single-stranded DNA template" *Methods in Enzymol.* 154:329-350); phosphorothioate-modified DNA mutagenesis (Taylor et al. (1985) "The use of phosphorothioate-modified DNA in restriction enzyme reactions to prepare nicked DNA" *Nucl. Acids Res.* 13: 8749-8764; Taylor et al. (1985) "The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA" *Nucl. Acids Res.* 13: 8765-8787 (1985); Nakamaye & Eckstein (1986) "Inhibition of restriction endonuclease Nci I cleavage by phosphorothioate groups and its application to oligonucleotide-directed mutagenesis" *Nucl. Acids Res.* 14: 9679-9698; Sayers et al. (1988) "Y-T Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis" *Nucl. Acids Res.* 16:791-802; and Sayers et al. (1988) "Strand specific cleavage of phosphorothioate-containing DNA by reaction with restriction endonucleases in the presence of ethidium bromide" *Nucl. Acids Res.* 16: 803-814); mutagenesis using gapped duplex DNA (Kramer et al. (1984) "The gapped duplex DNA approach to oligonucleotide-directed mutation construction" *Nucl. Acids Res.* 12: 9441-9456; Kramer & Fritz (1987) *Methods in Enzymol.*

"Oligonucleotide-directed construction of mutations via gapped duplex DNA" 154:350-367; Kramer et al. (1988) "Improved enzymatic in vitro reactions in the gapped duplex DNA approach to oligonucleotide-directed construction of mutations" *Nucl. Acids Res.* 16: 7207; and Fritz et al. (1988) "Oligonucleotide-directed construction of mutations: a gapped duplex DNA procedure without enzymatic reactions in vitro" *Nucl. Acids Res.* 16: 6987-6999).

Additional suitable methods include point mismatch repair (Kramer et al. (1984) "Point Mismatch Repair" *Cell* 38:879-887), mutagenesis using repair-deficient host strains (Carter et al. (1985) "Improved oligonucleotide site-directed mutagenesis using M13 vectors" *Nucl. Acids Res.* 13: 4431-4443; and Carter (1987) "Improved oligonucleotide-directed mutagenesis using M13 vectors" *Methods in Enzymol.* 154: 382-403), deletion mutagenesis (Eghtedarzadeh & Henikoff (1986) "Use of oligonucleotides to generate large deletions" *Nucl. Acids Res.* 14: 5115), restriction-selection and restriction-selection and restriction-purification (Wells et al. (1986) "Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin" *Phil. Trans. R. Soc. Lond.* A 317: 415-423), mutagenesis by total gene synthesis (Nambiar et al. (1984) "Total synthesis and cloning of a gene coding for the ribonuclease S protein" *Science* 223: 1299-1301; Sakamar and Khorana (1988) "Total synthesis and expression of a gene for the a-subunit of bovine rod outer segment guanine nucleotide-binding protein (transducin)" *Nucl. Acids Res.* 14: 6361-6372; Wells et al. (1985) "Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites" *Gene* 34:315-323; and Grundström et al. (1985) "Oligonucleotide-directed mutagenesis by microscale 'shot-gun' gene synthesis" *Nucl. Acids Res.* 13: 3305-3316), double-strand break repair (Mandecki (1986); Arnold (1993) "Protein engineering for unusual environments" *Current Opinion in Biotechnology* 4:450-455. "Oligonucleotide-directed double-strand break repair in plasmids of *Escherichia coli*: a method for site-specific mutagenesis" *Proc. Natl. Acad. Sci. USA*, 83:7177-7181). Additional details on many of the above methods can be found in *Methods in Enzymology* Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

Additional details regarding various diversity generating methods can be found in the following U.S. patents, PCT publications, and EPO publications: U.S. Pat. No. 5,605,793 to Stemmer (February 25, 1997), "Methods for In Vitro

Recombination;" U.S. Pat. No. 5,811,238 to Stemmer et al. (September 22, 1998) "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" U.S. Pat. No. 5,830,721 to Stemmer et al. (November 3, 1998), "DNA Mutagenesis by Random Fragmentation and Reassembly;" U.S. Pat. No. 5,834,252 to Stemmer, et al. (November 10, 1998) "End-Complementary Polymerase Reaction;" U.S. Pat. No. 5,837,458 to Minshull, et al. (November 17, 1998), "Methods and Compositions for Cellular and Metabolic Engineering;" WO 95/22625, Stemmer and Crameri, "Mutagenesis by Random Fragmentation and Reassembly;" WO 96/33207 by Stemmer and Lipschutz "End Complementary Polymerase Chain Reaction;" WO 97/20078 by Stemmer and Crameri "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" WO 97/35966 by Minshull and Stemmer, "Methods and Compositions for Cellular and Metabolic Engineering;" WO 99/41402 by Punnonen et al. "Targeting of Genetic Vaccine Vectors;" WO 99/41383 by Punnonen et al. "Antigen Library Immunization;" WO 99/41369 by Punnonen et al. "Genetic Vaccine Vector Engineering;" WO 99/41368 by Punnonen et al. "Optimization of Immunomodulatory Properties of Genetic Vaccines;" EP 752008 by Stemmer and Crameri, "DNA Mutagenesis by Random Fragmentation and Reassembly;" EP 0932670 by Stemmer "Evolving Cellular DNA Uptake by Recursive Sequence Recombination;" WO 99/23107 by Stemmer et al., "Modification of Virus Tropism and Host Range by Viral Genome Shuffling;" WO 99/21979 by Apt et al., "Human Papillomavirus Vectors;" WO 98/31837 by del Cardayre et al. "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination;" WO 98/27230 by Patten and Stemmer, "Methods and Compositions for Polypeptide Engineering;" WO 98/27230 by Stemmer et al., "Methods for Optimization of Gene Therapy by Recursive Sequence Shuffling and Selection," WO 00/00632, "Methods for Generating Highly Diverse Libraries," WO 00/09679, "Methods for Obtaining in Vitro Recombined Polynucleotide Sequence Banks and Resulting Sequences," WO 98/42832 by Arnold et al., "Recombination of Polynucleotide Sequences Using Random or Defined Primers," WO 99/29902 by Arnold et al., "Method for Creating Polynucleotide and Polypeptide Sequences," WO 98/41653 by Vind, "An in Vitro Method for Construction of a DNA Library," WO 98/41622 by Borchert et al.,

"Method for Constructing a Library Using DNA Shuffling," and WO 98/42727 by Pati and Zarling, "Sequence Alterations using Homologous Recombination."

Certain U.S. applications provide additional details regarding various diversity generating methods, including "SHUFFLING OF CODON ALTERED GENES" by Patten et al. filed September 28, 1999, (USSN 09/407,800); "EVOLUTION OF WHOLE CELLS AND ORGANISMS BY RECURSIVE SEQUENCE RECOMBINATION", by del Cardayre et al. filed July 15, 1998 (USSN 09/166,188), and July 15, 1999 (USSN 09/354,922); "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed September 28, 1999 (USSN 09/408,392), and "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed January 18, 2000 (PCT/US00/01203); "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed January 18, 2000, (PCT/US00/01202) and, e.g., "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed July 18, 2000 (USSN 09/618,579); "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLUTIONARY SIMULATIONS" by Selifonov and Stemmer, filed January 18, 2000 (PCT/US00/01138); and "SINGLE-STRANDED NUCLEIC ACID TEMPLATE-MEDIATED RECOMBINATION AND NUCLEIC ACID FRAGMENT ISOLATION" by Affholter, USSN 09/656,549, filed Sept. 6, 2000.

In brief, several different general classes of sequence modification methods, such as mutation, recombination, etc. are applicable to the present invention and set forth, e.g., in the references above. The following exemplify some of the different types of preferred formats for diversity generation that are optionally adapted to the present invention to create further diversity in, e.g., the recombined nucleic acid or gene sequences, or in the substrates for recombination (e.g., codon-varied oligonucleotides, fragmented nucleic acids, or the like) discussed herein, to produce new proteins or other expression products with improved properties.

Nucleic acids can be recombined in vitro by any of a variety of techniques discussed in the references above, including e.g., DNAse digestion of nucleic acids to be recombined followed by ligation and/or PCR reassembly of the

nucleic acids. In vitro recombination is described further, *supra*. For example, sexual PCR mutagenesis can be used in which random (or pseudo random, or even non-random) fragmentation of the DNA molecule is followed by recombination, based on sequence similarity, between DNA molecules with different but related DNA sequences, in vitro, followed by fixation of the crossover by extension in a polymerase chain reaction. This process and many process variants is described in several of the references above, e.g., in Stemmer (1994) *Proc. Natl. Acad. Sci. USA* 91:10747-10751. Any such recombination reaction can be spiked with a codon-varied oligonucleotide made according to the methods described herein.

Similarly, nucleic acids can be recursively recombined in vivo, e.g., by allowing recombination to occur between nucleic acids in cells. In vivo recombination is described further, *supra*. Many such in vivo recombination formats are set forth in the references noted above. Such formats optionally provide direct recombination between nucleic acids of interest, or provide recombination between vectors, viruses, plasmids, etc., comprising the nucleic acids of interest, as well as other formats. Details regarding such procedures are found in the references noted above.

Whole genome recombination methods can also be used in which whole genomes of cells or other organisms are recombined, optionally including spiking of the genomic recombination mixtures with desired library components (e.g., genes corresponding to the pathways of the present invention). These methods have many applications, including those in which the identity of a target gene is not known. Details on such methods are found, e.g., in WO 98/31837 by del Cardayre et al. "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination;" and in, e.g., PCT/US99/15972 by del Cardayre et al., also entitled "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination."

Synthetic recombination methods can also be used, in which oligonucleotides corresponding to targets of interest are synthesized and reassembled in PCR or ligation reactions which include oligonucleotides which correspond to more than one parental nucleic acid, thereby generating new recombined nucleic acids. Oligonucleotides can be made by standard nucleotide addition methods, or can be made, e.g., by tri-nucleotide synthetic approaches described herein. Details regarding such approaches are found in the references noted above, including, e.g.,

"OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed September 28, 1999 (USSN 09/408,392), and "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., filed January 18, 2000 (PCT/US00/01203); "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed January 18, 2000, (PCT/US00/01202); "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLUTIONARY SIMULATIONS" by Selifonov and Stemmer (PCT/US00/01138), filed January 18, 2000; and, e.g., "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed July 18, 2000 (USSN 09/618,579).

Many methods of accessing natural diversity, e.g., by hybridization of diverse nucleic acids or nucleic acid fragments to single-stranded templates, followed by polymerization and/or ligation to regenerate full-length sequences, optionally followed by degradation of the templates and recovery of the resulting modified nucleic acids can be similarly used. In one method employing a single-stranded template, the fragment population derived from the genomic library(ies) is annealed with partial, or, often approximately full length ssDNA or RNA corresponding to the opposite strand. Assembly of complex chimeric genes from this population is then mediated by nuclease-base removal of non-hybridizing fragment ends, polymerization to fill gaps between such fragments and subsequent single stranded ligation. The parental polynucleotide strand can be removed by digestion (e.g., if RNA or uracil-containing), magnetic separation under denaturing conditions (if labeled in a manner conducive to such separation) and other available separation/purification methods. Alternatively, the parental strand is optionally co-purified with the chimeric strands and removed during subsequent screening and processing steps. Additional details regarding this approach are found, e.g., in "SINGLE-STRANDED NUCLEIC ACID TEMPLATE-MEDIATED RECOMBINATION AND NUCLEIC ACID FRAGMENT ISOLATION" by Affholter, USSN 09/656,549, filed Sept. 6, 2000.

In silico methods of recombination can be effected in which genetic algorithms are used in a computer to recombine sequence strings which correspond to homologous (or even non-homologous) nucleic acids. The resulting recombined

sequence strings are optionally converted into nucleic acids by synthesis of nucleic acids which correspond to the recombined sequences, e.g., in concert with oligonucleotide synthesis/ gene reassembly techniques. This approach can generate random, partially random or designed variants. Many details regarding in silico recombination, including the use of genetic algorithms, genetic operators and the like in computer systems, combined with generation of corresponding nucleic acids (and/or proteins), as well as combinations of designed nucleic acids and/or proteins (e.g., based on cross-over site selection) as well as designed, pseudo-random or random recombination methods are described in "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed January 18, 2000, (PCT/US00/01202) "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLUTIONARY SIMULATIONS" by Selifonov and Stemmer (PCT/US00/01138), filed January 18, 2000; and, e.g., "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed July 18, 2000 (USSN 09/618,579). Extensive details regarding in silico recombination methods are found in these applications. This methodology is generally applicable to the present invention in providing, e.g., for codon-varied oligonucleotide-mediated recombination in silico and/or the generation of corresponding nucleic acids or proteins.

In another approach, single-stranded molecules are converted to double-stranded DNA (dsDNA) and the dsDNA molecules are bound to a solid support by ligand-mediated binding. After separation of unbound DNA, the selected DNA molecules are released from the support and introduced into a suitable host cell to generate a library enriched sequences which hybridize to the probe. A library produced in this manner provides a desirable substrate for further diversification using any of the procedures described herein.

Any of the preceding general recombination formats can be practiced in a reiterative fashion (e.g., one or more cycles of mutation/recombination or other diversity generation methods, optionally followed by one or more selection methods) to generate a more diverse set of recombinant nucleic acids.

Mutagenesis employing polynucleotide chain termination methods have also been proposed (*see* e.g., U.S. Patent No. 5,965,408, "Method of DNA

reassembly by interrupting synthesis" to Short, and the references above), and can be applied to the present invention. In this approach, double stranded DNAs corresponding to one or more genes sharing regions of sequence similarity are combined and denatured, in the presence or absence of primers specific for the gene. The single stranded polynucleotides are then annealed and incubated in the presence of a polymerase and a chain terminating reagent (e.g., ultraviolet, gamma or X-ray irradiation; ethidium bromide or other intercalators; DNA binding proteins, such as single strand binding proteins, transcription activating factors, or histones; polycyclic aromatic hydrocarbons; trivalent chromium or a trivalent chromium salt; or abbreviated polymerization mediated by rapid thermocycling; and the like), resulting in the production of partial duplex molecules. The partial duplex molecules, e.g., containing partially extended chains, are then denatured and reannealed in subsequent rounds of replication or partial replication resulting in polynucleotides which share varying degrees of sequence similarity and which are diversified with respect to the starting population of DNA molecules. Optionally, the products, or partial pools of the products, can be amplified at one or more stages in the process. Polynucleotides produced by a chain termination method, such as described above, are suitable substrates for any other described recombination format.

Diversity also can be generated in nucleic acids or populations of nucleic acids using a recombinational procedure termed "incremental truncation for the creation of hybrid enzymes" ("ITCHY") described in Ostermeier et al. (1999) "A combinatorial approach to hybrid enzymes independent of DNA homology" *Nature Biotech* 17:1205. This approach can be used to generate an initial a library of variants which can optionally serve as a substrate for one or more in vitro or in vivo recombination methods. See, also, Ostermeier et al. (1999) "Combinatorial Protein Engineering by Incremental Truncation," *Proc. Natl. Acad. Sci. USA*, 96: 3562-67; Ostermeier et al. (1999), "Incremental Truncation as a Strategy in the Engineering of Novel Biocatalysts," *Biological and Medicinal Chemistry*, 7: 2139-44.

Mutational methods which result in the alteration of individual nucleotides or groups of contiguous or non-contiguous nucleotides can be favorably employed to introduce nucleotide diversity. Many mutagenesis methods are found in the above-cited references; additional details regarding mutagenesis methods can be found in the following, which can also be applied to the present invention.

For example, error-prone PCR can be used to generate nucleic acid variants. Using this technique, PCR is performed under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. Examples of such techniques are found in the references above and, e.g., in Leung et al. (1989) *Technique* 1:11-15 and Caldwell et al. (1992) *PCR Methods Applic.* 2:28-33. Similarly, assembly PCR can be used, in a process which involves the assembly of a PCR product from a mixture of small DNA fragments. A large number of different PCR reactions can occur in parallel in the same reaction mixture, with the products of one reaction priming the products of another reaction.

Oligonucleotide directed mutagenesis can be used to introduce site-specific mutations in a nucleic acid sequence of interest. Examples of such techniques are found in the references above and, e.g., in Reidhaar-Olson et al. (1988) *Science*, 241:53-57. Similarly, cassette mutagenesis can be used in a process that replaces a small region of a double stranded DNA molecule with a synthetic oligonucleotide cassette that differs from the native sequence. The oligonucleotide can contain, e.g., completely and/or partially randomized native sequence(s).

Recursive ensemble mutagenesis is a process in which an algorithm for protein mutagenesis is used to produce diverse populations of phenotypically related mutants, members of which differ in amino acid sequence. This method uses a feedback mechanism to monitor successive rounds of combinatorial cassette mutagenesis. Examples of this approach are found in Arkin & Youvan (1992) *Proc. Natl. Acad. Sci. USA* 89:7811-7815.

Exponential ensemble mutagenesis can be used for generating combinatorial libraries with a high percentage of unique and functional mutants. Small groups of residues in a sequence of interest are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. Examples of such procedures are found in Delegrave & Youvan (1993) *Biotechnology Research* 11:1548-1552.

In vivo mutagenesis can be used to generate random mutations in any cloned DNA of interest by propagating the DNA, e.g., in a strain of *E. coli* that carries mutations in one or more of the DNA repair pathways. These "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA

in one of these strains will eventually generate random mutations within the DNA. Such procedures are described in the references noted above.

Other procedures for introducing diversity into a genome, e.g. a bacterial, fungal, animal or plant genome can be used in conjunction with the above described and/or referenced methods. For example, in addition to the methods above, techniques have been proposed which produce nucleic acid multimers suitable for transformation into a variety of species (*see*, e.g., Schellenberger U.S. Patent No. 5,756,316 and the references above). Transformation of a suitable host with such multimers, consisting of genes that are divergent with respect to one another, (e.g., derived from natural diversity or through application of site directed mutagenesis, error prone PCR, passage through mutagenic bacterial strains, and the like), provides a source of nucleic acid diversity for DNA diversification, e.g., by an in vivo recombination process as indicated above.

Alternatively, a multiplicity of monomeric polynucleotides sharing regions of partial sequence similarity can be transformed into a host species and recombined in vivo by the host cell. Subsequent rounds of cell division can be used to generate libraries, members of which, include a single, homogenous population, or pool of monomeric polynucleotides. Alternatively, the monomeric nucleic acid can be recovered by standard techniques, e.g., PCR and/or cloning, and recombined in any of the recombination formats, including recursive recombination formats, described above.

Methods for generating multispecies expression libraries have been described (in addition to the reference noted above, *see*, e.g., Peterson et al. (1998) U.S. Pat. No. 5,783,431 "METHODS FOR GENERATING AND SCREENING NOVEL METABOLIC PATHWAYS," and Thompson, et al. (1998) U.S. Pat. No. 5,824,485 METHODS FOR GENERATING AND SCREENING NOVEL METABOLIC PATHWAYS) and their use to identify protein activities of interest has been proposed (In addition to the references noted above, *see*, Short (1999) U.S. Pat. No. 5,958,672 "PROTEIN ACTIVITY SCREENING OF CLONES HAVING DNA FROM UNCULTIVATED MICROORGANISMS"). Multispecies expression libraries include, in general, libraries comprising cDNA or genomic sequences from a plurality of species or strains, operably linked to appropriate regulatory sequences, in an expression cassette. The cDNA and/or genomic sequences are optionally

randomly ligated to further enhance diversity. The vector can be a shuttle vector suitable for transformation and expression in more than one species of host organism, e.g., bacterial species, eukaryotic cells. In some cases, the library is biased by preselecting sequences which encode a protein of interest, or which hybridize to a nucleic acid of interest. Any such libraries can be provided as substrates for any of the methods herein described.

The above descibed procedures have been largely directed to increasing nucleic acid and/ or encoded protein diversity. However, in many cases, not all of the diversity is useful, e.g., functional, and contributes merely to increasing the background of variants that must be screened or selected to identify the few favorable variants. In some applications, it is desirable to preselect or prescreen libraries (e.g., an amplified library, a genomic library, a cDNA library, a normalized library, etc.) or other substrate nucleic acids prior to diversification, e.g., by recombination-based mutagenesis procedures, or to otherwise bias the substrates towards nucleic acids that encode functional products. For example, in the case of antibody engineering, it is possible to bias the diversity generating process toward antibodies with functional antigen binding sites by taking advantage of in vivo recombination events prior to manipulation by any of the described methods. For example, recombined CDRs derived from B cell cDNA libraries can be amplified and assembled into framework regions (e.g., Jirholt et al. (1998) "Exploiting sequence space: shuffling in vivo formed complementarity determining regions into a master framework" *Gene* 215: 471) prior to diversifying according to any of the methods described herein.

Libraries can be biased towards nucleic acids which encode proteins with desirable enzyme activities. For example, after identifying a clone from a library which exhibits a specified activity, the clone can be mutagenized using any known method for introducing DNA alterations. A library comprising the mutagenized homologues is then screened for a desired activity, which can be the same as or different from the initially specified activity. An example of such a procedure is proposed in Short (1999) U.S. Patent No. 5,939,250 for "PRODUCTION OF ENZYMES HAVING DESIRED ACTIVITIES BY MUTAGENESIS." Desired activities can be identified by any method known in the art. For example, WO 99/10539 proposes that gene libraries can be screened by combining extracts from the

gene library with components obtained from metabolically rich cells and identifying combinations which exhibit the desired activity. It has also been proposed (e.g., WO 98/58085) that clones with desired activities can be identified by inserting bioactive substrates into samples of the library, and detecting bioactive fluorescence corresponding to the product of a desired activity using a fluorescent analyzer, e.g., a flow cytometry device, a CCD, a fluorometer, or a spectrophotometer.

Libraries can also be biased towards nucleic acids which have specified characteristics, e.g., hybridization to a selected nucleic acid probe. For example, application WO 99/10539 proposes that polynucleotides encoding a desired activity (e.g., an enzymatic activity, for example: a lipase, an esterase, a protease, a glycosidase, a glycosyl transferase, a phosphatase, a kinase, an oxygenase, a peroxidase, a hydrolase, a hydratase, a nitrilase, a transaminase, an amidase or an acylase) can be identified from among genomic DNA sequences in the following manner. Single stranded DNA molecules from a population of genomic DNA are hybridized to a ligand-conjugated probe. The genomic DNA can be derived from either a cultivated or uncultivated microorganism, or from an environmental sample. Alternatively, the genomic DNA can be derived from a multicellular organism, or a tissue derived therefrom. Second strand synthesis can be conducted directly from the hybridization probe used in the capture, with or without prior release from the capture medium or by a wide variety of other strategies known in the art. Alternatively, the isolated single-stranded genomic DNA population can be fragmented without further cloning and used directly in, e.g., a recombination-based approach, that employs a single-stranded template, as described herein.

"Non-Stochastic" methods of generating nucleic acids and polypeptides are alleged in Short "Non-Stochastic Generation of Genetic Vaccines and Enzymes" WO 00/46344. These methods, including proposed non-stochastic polynucleotide reassembly and site-saturation mutagenesis methods can be applied to the present invention as well. Random or semi-random mutagenesis using doped or degenerate oligonucleotides is also described in, e.g., Arkin and Youvan (1992) "Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis" *Biotechnology* 10:297-300; Reidhaar-Olson et al. (1991) "Random mutagenesis of protein sequences using oligonucleotide cassettes" *Methods Enzymol.* 208:564-86; Lim and Sauer (1991) "The role of internal packing

interactions in determining the structure and stability of a protein" *J. Mol. Biol.* 219:359-76; Breyer and Sauer (1989) "Mutational analysis of the fine specificity of binding of monoclonal antibody 51F to lambda repressor" *J. Biol. Chem.* 264:13355-60); and "Walk-Through Mutagenesis" (Crea, R; US Patents 5,830,650 and 5,798,208, and EP Patent 0527809 B1.

It will readily be appreciated that any of the above described techniques suitable for enriching a library prior to diversification can also be used to screen the products, or libraries of products, produced by the diversity generating methods.

Kits for mutagenesis, library construction and other diversity generation methods are also commercially available. For example, kits are available from, e.g., Stratagene (e.g., QuickChange™ site-directed mutagenesis kit; and Chameleon™ double-stranded, site-directed mutagenesis kit), Bio/Can Scientific, Bio-Rad (e.g., using the Kunkel method described above), Boehringer Mannheim Corp., Clonetech Laboratories, DNA Technologies, Epicentre Technologies (e.g., 5 prime 3 prime kit); Genpak Inc, Lemargo Inc, Life Technologies (Gibco BRL), New England Biolabs, Pharmacia Biotech, Promega Corp., Quantum Biotechnologies, Amersham International plc (e.g., using the Eckstein method above), and Anglian Biotechnology Ltd (e.g., using the Carter/Winter method above).

The above references provide many mutational formats, including recombination, recursive recombination, recursive mutation and combinations or recombination with other forms of mutagenesis, as well as many modifications of these formats. Regardless of the diversity generation format that is used, the nucleic acids of the invention (e.g., codon-varied oligonucleotides, fragmented parental nucleic acids, or the like) can be recombined (with each other, or with related (or even unrelated) sequences) to produce a diverse set of recombinant nucleic acids, including, e.g., sets of homologous nucleic acids, as well as corresponding polypeptides. Any of the methods in the references above can be used in combination with any method herein, to provide substrates to the reactions noted herein, or to further modify the recombined nucleic acids produced according to the methods herein.

SELECTION OF A DESIRED TRAIT OR PROPERTY

The exact nature of the selection or screening method that is used following the recombination procedures herein is not a critical aspect of the invention. One or more recombination cycle(s) is/are optionally followed by at least one cycle of screening or selection for molecules having desired traits or properties. Various exemplary "breedable" traits or properties for which, e.g., evolved biocatalysts can be screened or selected include assorted kinetic constants, stability, selectivity, inhibition profiles, altered substrate specificity, increased enantioselectivity, increased activity, increased gene expression, activity under diverse environmental conditions (i.e., increased thermostability, increased activity in various organic solvents, pH tolerance, etc.), or the like. Generally, one or more recombination cycle(s) is/are optionally followed by at least one cycle of selection for molecules having one or more of these or other desired traits or properties. Many other desired traits or properties that are optionally screened or selected for according to the methods herein are described in greater detail in the references cited herein.

If a recombination cycle is performed *in vitro*, the products of recombination, *i.e.*, recombinant nucleic acids, are sometimes introduced into cells before the selection step. Recombinant nucleic acids can also be linked to an appropriate vector or to other regulatory sequences before selection. Alternatively, products of recombination generated *in vitro* are sometimes packaged in viruses (e.g., bacteriophage) before selection. If recombination is performed *in vivo*, recombination products may sometimes be selected in the cells in which recombination occurred. In other applications, recombinant segments are extracted from the cells, and optionally packaged as viruses or other vectors, before selection.

The nature of selection depends on what trait or property is to be acquired or for which improvement is sought. It is not usually necessary to understand the molecular basis by which particular recombination products have acquired new or improved traits or properties relative to the starting substrates. For instance, a gene has many component sequences, each having a different intended role (*e.g.*, coding sequences, regulatory sequences, targeting sequences, stability-conferring sequences, subunit sequences and sequences affecting integration). Each of these component sequences are optionally varied and recombined simultaneously. Selection is then performed, for example, for recombinant products that have an

increased ability to confer activity upon a cell without the need to attribute such improvement to any of the individual component sequences of the vector.

Depending on the particular protocol used to select for a desired trait or property, initial round(s) of screening can sometimes be performed using bacterial cells due to high transfection efficiencies and ease of culture. However, yeast, fungal or other eukaryotic systems may also be used for library expression and screening when bacterial expression is not practical or desired. Similarly, other types of selection that are not amenable to screening in bacterial or simple eukaryotic library cells, are performed in cells selected for use in an environment close to that of their intended use. Final rounds of screening are optionally performed in the precise cell type of intended use.

When further improvement in a trait is sought, at least one and usually a collection of recombinant products surviving a first round of screening/selection are optionally subject to a further round of recombination. These recombinant products can be recombined with each other or with exogenous segments representing the original substrates or further variants thereof. Again, recombination can proceed *in vitro* or *in vivo*. If the previous screening step identifies desired recombinant products as components of cells, the components can be subjected to further recombination *in vivo*, or can be subjected to further recombination *in vitro*, or can be isolated before performing a round of *in vitro* recombination. Conversely, if the previous selection step identifies desired recombinant products in naked form or as components of viruses, these segments can be introduced into cells to perform a round of *in vivo* recombination. The second round of recombination, irrespective how performed, generates additionally recombined products which encompass more diversity than is present in recombinant products resulting from previous rounds.

The second round of recombination may be followed by still further rounds of screening/selection according to the principles discussed for the first round. The stringency of selection can be increased between rounds. Also, the nature of the screen and the trait or property being selected may be varied between rounds if improvement in more than one trait or property is sought. Additional rounds of recombination and screening can then be performed until the recombinant products have sufficiently evolved to acquire the desired new or improved trait or property.

Multiple cycles of recombination can be performed to increase library diversity before a round of selection is performed. Alternately, where the library is diverse, multiple rounds of selection can be performed prior to recombination methods.

In the context of a particular experiment, a variety of related (or even unrelated) properties can be selected for using any available assay. For example, screening assays for an evolved dehalogenase activity can be performed, e.g., by detecting hydronium ions or halide ions liberated upon hydrolysis of, e.g., carbon-halogen bonds in reactant or substrate molecules. Other suitable techniques can include alcohol dehydrogenase-linked enzyme assays, fluorescence resonance energy transfer (FRET) assays, gas chromatography mass spectroscopy (GCMS) analysis, or the like. Furthermore, multiple traits or properties are optionally screened or selected for, e.g., sequentially or simultaneously.

General texts that describe molecular biological techniques useful herein, including mutagenesis, library construction, screening assays, cell culture and the like include Berger and Kimmel, *Guide to Molecular Cloning Techniques, Methods in Enzymology* volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook *et al.*, *Molecular Cloning - A Laboratory Manual* (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 (Sambrook); and *Current Protocols in Molecular Biology*, F.M. Ausubel *et al.*, eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1998) (Ausubel). Methods of transducing cells, including plant and animal cells, with nucleic acids are generally available, as are methods of expressing proteins encoded by such nucleic acids. In addition to Berger, Ausubel and Sambrook, useful general references for culture of animal cells include Freshney (*Culture of Animal Cells, a Manual of Basic Technique*, third edition Wiley-Liss, New York (1994)) and the references cited therein, Humason (*Animal Tissue Techniques*, fourth edition W.H. Freeman and Company (1979)) and Ricciardelli, *et al.*, *In Vitro Cell Dev. Biol.* 25:1016-1024 (1989). References for plant cell cloning, culture and regeneration include Payne et al. (1992) *Plant Cell and Tissue Culture in Liquid Systems* John Wiley & Sons, Inc. New York, NY (Payne); and Gamborg and Phillips (eds) (1995) *Plant Cell, Tissue and Organ Culture; Fundamental Methods* Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York) (Gamborg). A

variety of Cell culture media are described in Atlas and Parks (eds) *The Handbook of Microbiological Media* (1993) CRC Press, Boca Raton, FL (Atlas). Additional information for plant cell culture is found in available commercial literature such as the *Life Science Research Cell Culture Catalogue* (1998) from Sigma-Aldrich, Inc (St Louis, MO) (Sigma-LSRCCC) and, e.g., the *Plant Culture Catalogue* and supplement (1997) also from Sigma-Aldrich, Inc (St Louis, MO) (Sigma-PCCS).

COMPOSITION OF POPULATIONS TO BE RECOMBINED

The present invention provides for the recombination of codon-varied oligonucleotides (e.g., overlapping codon-varied oligonucleotides) that are derived from homologous or non-homologous nucleic acid sequences, or combinations of such sequences. As such, the conceivable compositions of populations to be recombined applicable to the present invention are infinite, and specific codon-varied oligonucleotide sequences to be recombined can be presented in biased or unbiased concentrations in the composition. Particular compositions that are relevant to the invention are discussed, *infra*.

The composition of the method of recombining at least two parental nucleic acids can be provided by initially aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity. Optionally, this initial step can include selecting non-homologous nucleic acids to be recombined. The next step, in the case of homologous sequence selection, includes synthesizing a plurality of codon-varied oligonucleotides corresponding to at least one region of sequence diversity. On the other hand, if non-homologous sequences are selected, the next step includes synthesizing a plurality of codon-varied oligonucleotides that correspond to at least one subsequence from each of the at least two parental nucleic acids. Then, irrespective of the sequence-type initially selected, at least one full-length nucleic acid that is identical to, or homologous with, at least one of the homologous, i.e., parental nucleic acid sequences is provided and fragmented by DNase cleavage. Finally, the resulting set of nucleic acid fragments is mixed with the plurality of codon-varied oligonucleotides to provide the composition comprising at least one set of fragmented parental nucleic acids corresponding to the at least two parental nucleic acids with the set of fragmented parental nucleic acids including a plurality of codon-varied oligonucleotides.

An additional aspect of the present invention is a composition including a library of codon-varied oligonucleotides that comprises a plurality of codon-varied oligonucleotide member types which correspond to a plurality of subsequence regions of a plurality of members of a selected set of a plurality of homologous or non-homologous target sequences, which member types comprise a plurality of members with at least one region of similarity and at least one region of diversity. The region of diversity includes at least one codon difference. Also, the plurality of oligonucleotide member types of this composition can be present in non-equimolar amounts.

The composition described above can include a plurality of subsequence regions that include a plurality of non-overlapping sequence regions of the selected set of a plurality of homologous target sequences. This composition can, alternatively, include a plurality of oligonucleotide member types, each having a sequence identical to at least one subsequence from at least one of the selected set of target sequences in which those target sequences are identical. The composition can also be a plurality of oligonucleotide member types comprising a plurality of homologous oligonucleotides corresponding to a homologous region from the selected set of a plurality of homologous target sequences, where each of the plurality of homologous oligonucleotides comprise at least one codon-varied subsequence.

The shuffling compositions of the invention can include at least one of: a polymerase, a thermostable DNA polymerase, a nucleic acid synthesis reagent, a buffer, a salt, magnesium, and at least one nucleic acid sequence comprising at least one of the plurality of members of the selected set of homologous target sequences. This composition can also include a plurality of oligonucleotide member types that is selected by aligning a plurality of homologous or non-homologous target sequences, determining at least one region of identity and at least one region of diversity and synthesizing the oligonucleotides to encode at least a portion of the at least one region of identity, or at least a portion of the at least one region of diversity, or at least a portion of both.

Additionally, the compositions can include a plurality of oligonucleotide member types comprising at least one member type comprising at least one sequence diversity domain, or a plurality of sequence diversity domains. In the case when the composition is a plurality of oligonucleotide member types

comprising at least one member type comprising a plurality of sequence diversity domains, the plurality of sequence diversity domains can correspond to adjacent sequence regions on a plurality of the plurality of homologous nucleic acids when the homologous nucleic acids are aligned. Finally, the composition can be a library that comprises a set of cross-over codon-varied oligonucleotides with each oligonucleotide member of the set of cross-over codon-varied oligonucleotides including a plurality of sequence diversity domains corresponding to a plurality of homologous nucleic acids.

UTILIZATION OF OLIGONUCLEOTIDE-MEDIATED BLENDING TO TUNE RECOMBINATION

In certain embodiments of the invention, recombination is biased by supplying non-equimolar ratios of codon-varied oligonucleotides to the recombination composition. In this aspect, unlike certain other methods provided by the present invention, equimolar ratios of codon-varied oligonucleotides in a set of such oligonucleotides to be recombined are *not* used to produce a library of recombinant nucleic acids. Instead, ratios of particular oligonucleotides which correspond to the sequences of a selected member or selected set of members of the nucleic acids from which the codon-varied oligonucleotides are derived are selected by the practicioner. Non-equimolar ratios of codon-varied oligonucleotides may be achieved by, e.g., synthesizing disproportionate amounts of the relevant codon-varied oligonucleotides and/or providing disproportionate amounts to the composition of nucleic acids to be recombined.

The general strategy of tuning recombination by selecting oligonucleotide proportions is applicable to recombination of any two nucleic acids, whether of high or low sequence similarity. However, one advantage of this method when compared to other gene recombination approaches is that the overall sequence identity of two sequences to be blended can be lower. Further, sometimes only selected regions are recombined, making it possible to take available structural or functional data into account in specifying how the blended gene is constructed. As such, sequence space which cannot be created by other recombination protocols is accessed by the blended gene and a higher percentage of active clones may be obtained if structural information is taken into consideration.

CODON-VARIED OLIGONUCLEOTIDE RECOMBINATION TARGETS

Virtually any nucleic acid can be recombined by the methods described in this disclosure. As noted above, common sequence repositories for known proteins include GenBank®, Entrez®, EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet.

One class of preferred targets for activation includes nucleic acids encoding therapeutic proteins, e.g., erythropoietin (EPO), insulin, peptide hormones, e.g., human growth hormone; growth factors and cytokines, e.g., epithelial Neutrophil Activating Peptide-78, GROα/MGSA, GROβ, GROγ, MIP-1α, MIP-1β, MCP-1, epidermal growth factor, fibroblast growth factor, hepatocyte growth factor, insulin-like growth factor, the interferons, the interleukins, keratinocyte growth factor, leukemia inhibitory factor, oncostatin M, PD-ECSF, PDGF, pleiotropin, SCF, c-kit ligand, VEGEF, G-CSF etc. Many of these proteins are commercially available (*see*, e.g., the Sigma-Aldrich Co. 1999 *Biochemicals and Reagents* catalogue and price list), and the corresponding genes are well-known.

Another class of preferred targets are transcriptional and expression activators. Example transcriptional and expression activators include genes and proteins that modulate cell growth, differentiation, regulation, or the like. Expression and transcriptional activators are found in prokaryotes, viruses, and eukaryotes, including fungi, plants, and animals, including mammals, providing a wide range of therapeutic targets. It will be appreciated that expression and transcriptional activators regulate transcription by many mechanisms, e.g., by binding to receptors, stimulating a signal transduction cascade, regulating expression of transcription factors, binding to promoters and enhancers, binding to proteins that bind to promoters and enhancers, unwinding DNA, splicing pre-mRNA, polyadenylating RNA, and degrading RNA. Expression activators include cytokines, inflammatory molecules, growth factors, their receptors, and oncogene products, e.g., interleukins (e.g., IL-1, IL-2, IL-8, etc.), interferons, FGF, IGF-I, IGF-II, FGF, PDGF, TNF, TGF-α, TGF-β, EGF, KGF, SCF/c-Kit, CD40L/CD40, VLA-4/VCAM-1, ICAM-1/LFA-1, and hyalurin/CD44; signal transduction molecules and corresponding oncogene products, e.g., Mos, Ras, Raf, and Met; and transcriptional activators and suppressors, e.g., p53, Tat, Fos, Myc, Jun, Myb, Rel, and steroid hormone receptors, e.g., those for

estrogen, progesterone, testosterone, aldosterone, the LDL receptor ligand and corticosterone.

Rnases, e.g., Onconase and EDN, are preferred targets for the synthetic methods herein, particularly those methods utilizing gene blending. One of skill will appreciate that both frog and human RNAses are known and are known to have a number of important pharmacological activities.

Similarly, proteins from infectious organisms for possible vaccine applications, described in more detail below, including infectious fungi, e.g., *Aspergillus*, *Candida* species; bacteria, particularly *E. coli*, which serves a model for pathogenic bacteria, as well as medically important bacteria such as *Staphylococci* (e.g., *aureus*), *Streptococci* (e.g., *pneumoniae*), *Clostridia* (e.g., *perfringens*), *Neisseria* (e.g., *gonorrhoea*), *Enterobacteriaceae* (e.g., *coli*), *Helicobacter* (e.g., *pylori*), *Vibrio* (e.g., *cholerae*), *Campylobacter* (e.g., *jejuni*), *Pseudomonas* (e.g., *aeruginosa*), *Haemophilus* (e.g., *influenzae*), *Bordetella* (e.g., *pertussis*), *Mycoplasma* (e.g., *pneumoniae*), *Ureaplasma* (e.g., *urealyticum*), *Legionella* (e.g., *pneumophilia*), *Spirochetes* (e.g., *Treponema*, *Leptospira*, and *Borrelia*), *Mycobacteria* (e.g., *tuberculosis*, *smegmatis*), *Actinomyces* (e.g., *israelii*), *Nocardia* (e.g., *asteroides*), *Chlamydia* (e.g., *trachomatis*), *Rickettsia*, *Coxiella*, *Ehrilichia*, *Rocholimaea*, *Brucella*, *Yersinia*, *Francisella*, and *Pasteurella*; protozoa such as sporozoa (e.g., *Plasmodia*), *rhizopods* (e.g., *Entamoeba*) and flagellates (*Trypanosoma*, *Leishmania*, *Trichomonas*, *Giardia*, etc.); viruses such as (+) RNA viruses (examples include Poxviruses e.g., *vaccinia*; Picornaviruses, e.g. *polio*; Togaviruses, e.g., *rubella*; Flaviviruses, e.g., HCV; and Coronaviruses), (-) RNA viruses (examples include Rhabdoviruses, e.g., VSV; Paramyxovimses, e.g., RSV; Orthomyxovimses, e.g., influenza; Bunyaviruses; and Arenaviruses), dsDNA viruses (Reoviruses, for example), RNA to DNA viruses, i.e., Retroviruses, e.g., especially HIV and HTLV, and certain DNA to RNA viruses, such as, Hepatitis B virus.

Other proteins relevant to non-medical uses, such as, inhibitors of transcription or toxins of crop pests, e.g., insects, fungi, weed plants, and the like, are also preferred targets for codon-varied oligonucleotide recombination. Industrially important enzymes, such as, monooxygenases (e.g., p450s), proteases, nucleases, and lipases are also preferred targets. As an example, subtilisin can be evolved by codon-varied oligonucleotides for homologous forms of the gene for subtilisin. Von der

Osten et al., *J. Biotechnol.* 28:55-68 (1993) provide an example subtilisin coding nucleic acid. Proteins which aid in folding such as the chaperonins are also preferred targets.

Preferred known genes suitable for codon-varied oligonucleotide-mediated recombination also include the following: Alpha-1 antitrypsin, Angiostatin, Antihemolytic factor, Apolipoprotein, Apoprotein, Atrial natriuretic factor, Atrial natriuretic polypeptide, Atrial peptides, C-X-C chemokines (e.g., T39765, NAP-2, ENA-78, Gro-a, Gro-b, Gro-c, IP-10, GCP-2, NAP-4, SDF-1, PF4, MIG), Calcitonin, CC chemokines (e.g., Monocyte chemoattractant protein-1, Monocyte chemoattractant protein-2, Monocyte chemoattractant protein-3, Monocyte inflammatory protein-1 alpha, Monocyte inflammatory protein-1 beta, RANTES, I309, R83915, R91733, HCC1, T58847, D31065, T64262), CD40 ligand, Collagen, Colony stimulating factor (CSF), Complement factor 5a, Complement inhibitor, Complement receptor 1, Factor IX, Factor VII, Factor VIII, Factor X, Fibrinogen, Fibronectin, Glucocerebrosidase, Gonadotropin, Hedgehog proteins (e.g., Sonic, Indian, Desert), Hemoglobin (for blood substitute; for radiosensitization), Hirudin, Human serum albumin, Lactoferrin, Luciferase, Neurturin, Neutrophil inhibitory factor (NIF), Osteogenic protein, Parathyroid hormone, Protein A, Protein G, Relaxin, Renin, Salmon calcitonin, Salmon growth hormone, Soluble complement receptor I, Soluble I-CAM 1, Soluble interleukin receptors (IL-1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15), Soluble TNF receptor, Somatomedin, Somatostatin, Somatotropin, Streptokinase, Superantigens, i.e., Staphylococcal enterotoxins (SEA, SEB, SEC1, SEC2, SEC3, SED, SEE), Toxic shock syndrome toxin (TSST-1), Exfoliating toxins A and B, Pyrogenic exotoxins A, B, and C, and M arthritides mitogen, Superoxide dismutase, Thymosin alpha 1, Tissue plasminogen activator, Tumor necrosis factor beta (TNF beta), Tumor necrosis factor receptor (TNFR), Tumor necrosis factor-alpha (TNF alpha) and Urokinase.

Small proteins, such as, defensins (antifungal proteins of about 50 amino acids, EF40 (an anti fungal protein of 28 amino acids), peptide antibiotics, and peptide insecticidal proteins are also preferred targets and exist as families of related proteins. Nucleic acids encoding small proteins are particularly preferred targets, because conventional recombination methods provide only limited product sequence diversity. This is because conventional recombination methodology produces

crossovers between homologous sequences about every 50-100 base pairs. This means that for very short recombination targets, cross-overs occur by standard techniques about once per molecule. In contrast, the codon-varied oligonucleotide recombination formats herein provide for recombination of small nucleic acids, as the practicioner selects any "cross-over" desired.

A variety of additional targets which can be modified according to the present invention are described, e.g., in "SINGLE-STRANDED NUCLEIC ACID TEMPLATE-MEDIATED RECOMBINATION AND NUCLEIC ACID FRAGMENT ISOLATION" by Affholter, USSN 09/656,549, filed Sept. 6, 2000.

SYSTEM INTEGRATION

As noted, *supra*, the initial codon-varied oligonucleotide sequence selection step of the invention can involve the alignment of nucleic acids using a computer and sequence alignment software. Other important integrated system components, however, can also provide for high-throughput screening assays, in addition to the coupling of such assays to oligonucleotide selection, synthesis and recombination (e.g., *in silico* recombination). Relevant assays will, naturally, depend on the application. There are many known assays for, e.g., proteins, receptors, and ligands. Formats include binding to immobilized components, cell or organismal viability, production of reporter compositions, and the like.

In one aspect, the computer system is used to perform *in silico* recombination of character strings that correspond to, e.g., codon-varied oligonucleotides. A variety of such methods are set forth in "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov and Stemmer, filed October 12, 1999 (USSN 09/416,375). In brief, genetic operators are used in genetic algorithms to change given sequences, e.g., by mimicking genetic events such as mutation, recombination, death and the like. Multi-dimensional analysis to optimize sequences can also be performed in the computer system, e.g., as described in the '375 application.

A digital system can also instruct an oligonucleotide synthesizer to synthesize codon-varied oligonucleotides, e.g., used for recombination according to the methods described herein, or to order those sequences from commercial sources

(e.g., by printing appropriate order forms or by linking to an order form on the internet).

The digital system can also include output elements for controlling nucleic acid synthesis (*e.g.*, based upon a sequence or an alignment of nucleic acid sequences as herein), i.e., an integrated system of the invention optionally includes an oligonucleotide synthesizer or an oligonucleotide synthesis controller for synthesizing, e.g., codon-varied oligonucleotides or other nucleic acid fragment sequences. For example, the synthesizer can be configured for trinucleotide synthetic methods as described herein. The system can also include other operations which occur downstream from an alignment or other operation performed using a character string corresponding to a sequence described herein.

In the high throughput assays of the invention, it is possible to screen up to several thousand different recombination products in a single day. For example, each well of a microtiter plate can be used to run a separate assay, or, if concentration or incubation time effects are to be observed, every 5-10 wells can test a single product. Thus, a single standard microtiter plate can assay about 100 (e.g., 96) reactions. If 1536 well plates are used, then a single plate can easily assay from about 100 to approximately 1500 different reactions. It is possible to assay several different plates per day; assay screens for up to about 6,000-20,000 different assays (i.e., involving different nucleic acids, encoded proteins, concentrations, etc.) is possible using the integrated systems of the invention. More recently, microfluidic approaches to reagent manipulation have been developed, e.g., by Caliper Technologies (Mountain View, CA).

A number of well known robotic systems have also been developed for solution phase chemistries useful in assay systems. These systems include automated workstations like the automated synthesis apparatus developed by Takeda Chemical Industries, LTD. (Osaka, Japan) and many robotic systems utilizing robotic arms (Zymate II, Zymark Corporation, Hopkinton, Mass.; Orca, Hewlett-Packard, Palo Alto, Calif.) which mimic the manual synthetic operations performed by a scientist. Any of the above devices are suitable for use with the present invention, e.g., for high-throughput screening of molecules assembled from the various oligonucleotide sets described herein. The nature and implementation of modifications to these

devices (if any) so that they can operate as discussed herein with reference to the integrated system will be apparent to persons skilled in the relevant art.

High throughput screening systems are commercially available (*see, e.g.*, Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, *etc.*). These systems typically automate entire procedures including all sample and reagent pipetting, liquid dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization. The manufacturers of such systems provide detailed protocols the various high throughput. Thus, for example, Zymark Corp. provides technical bulletins describing screening systems for detecting the modulation of gene transcription, ligand binding, and the like.

Optical images viewed (and, optionally, recorded) by a camera or other recording device (*e.g.*, a photodiode and data storage device) are optionally further processed in any of the embodiments herein, *e.g.*, by digitizing the image and/or storing and analyzing the image on a computer. A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical image, *e.g.*, using PC (Intel x86 or Pentium chip-compatible DOS™, OS2™ WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, or UNIX based (*e.g.*, SUN™ work station) computers. One conventional system carries light from the assay device to a cooled charge-coupled device (CCD) camera, in common use in the art. A CCD camera includes an array of picture elements (pixels). The light from the specimen is imaged on the CCD. Particular pixels corresponding to regions of the specimen (*e.g.*, individual hybridization sites on an array of biological polymers) are sampled to obtain light intensity readings for each position. Multiple pixels are processed in parallel to increase speed. The apparatus and methods of the invention are easily used for viewing any sample, e.g., by fluorescent or dark field microscopic techniques.

Integrated systems for assay analysis in the present invention optionally include a digital computer with high-throughput liquid control software, image analysis software, data interpretation software, a robotic liquid control armature for transferring codon-varied oligonucleotide solutions or codon-varied

oligonucleotide compositions from a source to a destination operably linked to the digital computer, an input device (e.g., a computer keyboard) for entering data to the digital computer to control high throughput liquid transfer by the robotic liquid control armature and, optionally, an image scanner for digitizing label signals from labeled assay component. The image scanner interfaces with the image analysis software to provide a measurement of probe label intensity.

These assay systems can also include integrated systems incorporating oligonucleotide selection elements, such as a computer, database with nucleic acid sequences of interest, sequence alignment software, and oligonucleotide selection software. Suitable alignment algorithms, e.g., BLAST and others are discussed, *supra*. However, sequence alignment can optionally be achieved manually. Once sequences to be synthesized are selected, they can be converted into lines of character string information in data sets in a computer corresponding to the desired codon-varied oligonucleotides to be obtained.

Additional software can be included, such as, components for ordering the selected oligonucleotides, and/or directing synthesis of oligonucleotides by an operably linked automated synthesizer. In this case, the character string information in the output of an integrated computer directs the robotic arm of the automated synthesizer to perform the steps necessary to synthesize the desired codon-varied oligonucleotide sequences.

Although, the integrated system elements of the invention optionally include any of the above components to facilitate high throughput recombination and selection. It will be appreciated that these high-throughput recombination elements can be in systems separate from those for performing selection assays, or as discussed, the two can be integrated.

Modifications can be made to the method and materials as hereinbefore described without departing from the spirit or scope of the invention as claimed, and the invention can be put to a number of different uses, including:

The use of an integrated system to select codon-varied oligonucleotides and to test recombined nucleic acids for activity, including in an iterative process.

An assay, kit or system utilizing a use of any one of the selection strategies, materials, components, methods or substrates hereinbefore described. Kits

will optionally additionally comprise instructions for performing methods or assays, packaging materials, one or more containers which contain assay, device or system components, or the like.

In an additional aspect, the present invention provides kits embodying the methods and apparatus herein. Kits of the invention optionally comprise one or more of the following: (1) a recombination component as described herein; (2) instructions for practicing the methods described herein, and/or for operating the codon-varied oligonucleotide synthesis or recombined nucleic acid selection procedures herein; (3) one or more assay component(s); (4) a container for holding nucleic acids or enzymes, other nucleic acids, transgenic plants, animals, cells, or the like and, (5) packaging materials.

In a further aspect, the present invention provides for the use of any component or kit herein, for the practice of any method or assay herein, and/or for the use of any apparatus or kit to practice any assay or method herein.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and apparatus described above may be used in various combinations. All publications, patents, patent applications, or other documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication, patent, patent application, or other document were individually indicated to be incorporated by reference for all purposes.

WHAT IS CLAIMED IS:

1. A method of recombining codon-varied oligonucleotides to provide a population of recombined nucleic acids, the method comprising:
(i) providing a set of nucleic acid fragments comprising a plurality of codon-varied oligonucleotides;
(ii) hybridizing the set of nucleic acid fragments; and,
(iii) elongating the set of hybridized nucleic acid fragments, thereby providing a population of recombined nucleic acids.

2. The method of claim 1, wherein the codon-varied oligonucleotides are overlapping.

3. The method of claim 1, wherein the codon-varied oligonucleotides are produced by trinucleotide synthesis.

4. The method of claim 1, wherein the codon-varied oligonucleotides are produced by automated trinucleotide synthesis.

5. The method of claim 1, wherein the codon-varied oligonucleotides are produced by split-pool trinucleotide synthesis.

6. The method of claim 1, wherein the providing step comprises:
(a) providing a solid-phase substrate sequence having a 5' terminus and at least one base, the 5' terminus and at least one base having protecting groups thereon;
(b) removing the 5' protecting group of the solid-phase substrate sequence to provide a 5' deprotected solid-phase substrate sequence;
(c) coupling the 5' deprotected solid-phase substrate sequence with a selected trinucleotide phosphoramidite sequence having a 3' terminus, a 5' terminus, and three base groups, the 3' terminus, the 5' prime terminus, and the three base groups having protecting groups thereon, thereby yielding an extended oligonucleotide sequence; and,
(d) repeating steps (b) and (c), wherein the extended oligonucleotide sequence yielded by each repeated step (c) becomes the solid-phase substrate sequence of the next repeated step (b) until a desired codon-varied oligonucleotide is obtained.

7. The method of claim 1, wherein the providing step comprises providing a solid-phase substrate sequence having a 5' terminus and at least one base, the 5' terminus and at least one base having protecting groups thereon, the solid-phase substrate sequence further comprising a 3' end that is covalently attached to a solid support.

8. The method of claim 1, wherein the providing step comprises coupling together one or more of: mononucleotides, trinucleotide phosphoramidite sequences, and oligonucleotides.

9. The method of claim 1, wherein the providing step is a split-pool synthesis format comprising:

(a) providing solid-phase substrate sequences, each having a 5' terminus and at least one base, the 5' terminus and at least one base having protecting groups thereon;

(b) removing the 5' protecting groups of the solid-phase substrate sequences to provide 5' deprotected solid-phase substrate sequences;

(c) coupling the 5' deprotected solid-phase substrate sequences with a population of a selected trinucleotide phosphoramidite sequence, each having a 3' terminus, a 5' terminus, and three base groups, the 3' terminus, the 5' terminus, and the three base groups having protecting groups thereon, thereby yielding extended oligonucleotide sequences;

(d) repeating steps (b) and (c), wherein the extended oligonucleotide sequences yielded by each step (c) become the solid-phase substrate sequences of the next step (b), until extended intermediate oligonucleotide sequences are produced;

(e) splitting the extended intermediate oligonucleotide sequences into two or more separate pools;

(f) removing the 5' protecting groups of the extended intermediate oligonucleotide sequences to provide 5' deprotected extended intermediate oligonucleotide sequences in the two or more separate pools;

(g) coupling the 5' deprotected extended intermediate oligonucleotide sequences with one or more selected mononucleotides, trinucleotide phosporamidite sequences, or oligonucleotides in the two or more separate pools, thereby yielding further extended intermediate oligonucleotide sequences;

(h) pooling the further extended intermediate oligonucleotide sequences from the two or more separate pools into a single pool; and,

(i) repeating steps (b) through (h), wherein the further extended intermediate oligonucleotide sequences in the single pool of each step (h) become the solid-phase substrate sequences of the next step (b), until desired codon-varied oligonucleotides are obtained.

10. The method of claim 1, wherein the set of nucleic acid fragments comprise a plurality of oligonucleotide member types which comprise consensus region subsequences derived from a plurality of homologous target nucleic acids.

11. The method of claim 1, wherein the set of nucleic acid fragments comprising a plurality of oligonucleotide member types comprise at least 3 member types.

12. The method of claim 1, wherein the set of nucleic acid fragments comprising a plurality of oligonucleotide member types comprise at least 5 member types.

13. The method of claim 1, wherein the set of nucleic acid fragments comprising a plurality of oligonucleotide member types comprise at least 10 member types.

14. The method of claim 1, wherein the set of nucleic acid fragments comprises a plurality of homologous oligonucleotide member types, wherein the homologous oligonucleotide member types are present in approximately equimolar amounts.

15. The method of claim 1, wherein the set of nucleic acid fragments comprises a plurality of homologous oligonucleotide member types, wherein the homologous oligonucleotide member types are present in approximately non-equimolar amounts.

16. The method of claim 1, wherein the hybridizing step occurs *in vitro*.

17. The method of claim 1, wherein the hybridizing step occurs *in vivo*.

18. The method of claim 1, the method further comprising:

denaturing the population of recombined nucleic acids, thereby providing denatured recombined nucleic acids;

re-hybridizing the denatured recombined nucleic acids;

elongating the resulting re-rehybridized recombined nucleic acids; and,

selecting at least one of the resulting elongated re-hybridized recombined nucleic acids for at least one desired trait or property.

19. The method of claim 1, further comprising selecting at least one member of the population of recombined nucleic acids for at least one desired trait or property.

20. The method of claim 1, further comprising selecting at least first and second nucleic acids to be recombined, wherein the set of nucleic acid fragments comprises a plurality of codon-varied nucleic acids which correspond to the first and second nucleic acids.

21. The method of claim 20, wherein the first and second nucleic acids are homologous.

22. The method of claim 21, wherein the method of recombining codon-varied oligonucleotides to provide a population of recombined nucleic acids is performed in an integrated system comprising a computer or computer readable medium comprising a data set corresponding to a set of character strings corresponding to the at least first and second nucleic acids and a sequence comparison instruction set for optimizing alignment of the at least first and second homologous nucleic acids.

23. The method of claim 20, wherein the first and second nucleic acids are non-homologous.

24. The method of claim 23, wherein the non-homologous first and second nucleic acids are synthesized using the split-pool synthesis format.

25. The method of claim 24, wherein the non-homologous first and second nucleic acids are less than 90 percent homologous.

26. The method of claim 25, wherein the split-pool synthesis format is module-based with a smallest module being a single trinucleotide in length and a larger module being at least 15 nucleotides in length.

27. The method of claim 23, wherein the non-homologous first and second nucleic acids are synthesized using the split-pool synthesis format performed in an automated synthesizer.

28. The method of claim 27, wherein the non-homologous first and second nucleic acids are less than 90 percent homologous.

29. The method of claim 28, wherein the split-pool synthesis format performed in an automated synthesizer is module-based with a smallest module being a single trinucleotide in length and a larger module being at least 15 nucleotides in length.

30. The method of claim 1, wherein the providing step is performed in an automated synthesizer which automatically performs the steps of:

(a) providing a solid-phase substrate sequence having a 5' terminus and at least one base, the 5' terminus and at least one base having protecting groups thereon;

(b) removing the 5' protecting group of the solid-phase substrate sequence to provide a 5' deprotected solid-phase substrate sequence;

(c) coupling the 5' deprotected solid-phase substrate sequence with a selected trinucleotide phosphoramidite sequence having a 3' terminus, a 5' terminus, and three base groups, the 3' terminus, the 5' prime terminus, and the three base groups having protecting groups thereon, thereby yielding an extended oligonucleotide sequence; and,

(d) repeating steps (b) and (c), wherein the extended oligonucleotide sequence yielded by each repeated step (c) becomes the solid-phase substrate sequence of the next step (b) until a desired codon-varied oligonucleotide is obtained.

31. The method of claim 30, the method further comprising inputting character string information into the automatic synthesizer corresponding to the desired codon-varied oligonucleotides to be obtained.

32. The method of claim 31, wherein the character string information corresponds to two or more nucleic acids to be recombined.

33. The method of claim 1, wherein the providing step is a split-pool synthesis format performed in an automated synthesizer which automatically performs the steps of:

(a) providing solid-phase substrate sequences having a 5' terminus and at least one base, the 5' terminus and at least one base having protecting groups thereon;

(b) removing the 5' protecting groups of the solid-phase substrate sequences to provide 5' deprotected solid-phase substrate sequences;

(c) coupling the 5' deprotected solid-phase substrate sequences with a population of a selected trinucleotide phosphoramidite sequence having a 3' terminus, a 5' terminus, and three base groups, the 3' terminus, the 5' terminus, and the three base groups having protecting groups thereon, thereby yielding extended oligonucleotide sequences;

(d) repeating steps (b) and (c), wherein the extended oligonucleotide sequences yielded by each step (c) become the solid-phase substrate sequences of the next step (b), until extended intermediate oligonucleotide sequences are provided;

(e) splitting the extended intermediate oligonucleotide sequences into two or more separate pools;

(f) removing the 5' protecting groups of the extended intermediate oligonucleotide sequences to provide 5' deprotected extended intermediate oligonucleotide sequences in the two or more separate pools;

(g) coupling the 5' deprotected extended intermediate oligonucleotide sequences with one or more selected mononucleotides, trinucleotide phosporamidite sequences, or oligonucleotides in the two or more separate pools, thereby yielding further extended intermediate oligonucleotide sequences;

(h) pooling the further extended intermediate oligonucleotide sequences from the two or more separate pools into a single pool; and,

(i) repeating steps (b) through (h), wherein the further extended intermediate oligonucleotide sequences in the single pool of each step (h) become the solid-phase substrate sequences of the next step (b), until desired codon-varied oligonucleotides are obtained.

34. The method of claim 33, wherein the providing step further comprises providing a solid-phase substrate sequence having a 3' terminus covalently attached to a solid support.

35. The method of claim 33, the method further comprising inputting character string information into the automatic synthesizer corresponding to the desired codon-varied oligonucleotides to be obtained.

36. The method of claim 35, wherein the character string information corresponds to two or more nucleic acids to be recombined.

37. The method of claim 1, the elongating step comprising elongating the set of hybridized nucleic acid fragments with a polymerase, a ligase, or both.

38. The method of claim 37, wherein the polymerase is a thermostable polymerase.

39. The method of claim 37, wherein the polymerase is a thermostable ligase.

40. The method of claim 1, the method further comprising the steps of:

denaturing the population of recombined nucleic acids, thereby providing denatured recombined nucleic acids;

re-hybridizing the denatured recombined nucleic acids;

elongating the resulting re-hybridized recombined nucleic acids; and,

repeating the denaturing, re-hybridizing and elongating steps at least once.

41. The method of claim 40, further comprising selecting at least one of the resulting re-hybridized recombined nucleic acids for at least one desired trait or property.

42. The method of claim 40, wherein a plurality of members of the population of recombined nucleic acids is selected for a desired trait or property, thereby providing first round selected nucleic acids, the method further comprising:

hybridizing an additional set of nucleic acid fragments, which additional set of nucleic acid fragments is derived from the first round selected nucleic acids; and,

elongating the hybridized additional set of nucleic acid fragments, thereby providing a population of further recombined nucleic acids.

43. The method of claim 42, further comprising sequencing the first round selected nucleic acids, wherein the additional set of nucleic acid fragments is derived from the first round selected nucleic acids by aligning sequences of the first round selected nucleic acids to identify regions of identity and regions of diversity in the first round selected nucleic acids, and synthesizing the additional set of nucleic acid fragments to comprise a plurality of codon-varied oligonucleotides, each of which comprise subsequences corresponding to at least one region of diversity.

44. The method of claim 42, wherein the first round selected nucleic acids encode polypeptides of about 50 amino acids or less.

45. The method of claim 42, wherein the additional set of nucleic acid fragments comprise a plurality of oligonucleotide member types which comprise consensus region subsequences derived from a plurality of the first round selected nucleic acids.

46. A method of recombining at least two parental nucleic acids to provide at least one recombinant nucleic acid, the method comprising:

providing a composition comprising at least one set of fragmented parental nucleic acids corresponding to the at least two parental nucleic acids, the set of fragmented parental nucleic acids comprising a plurality of codon-varied oligonucleotides;

hybridizing the composition to provide at least one hybridized nucleic acid; and,

elongating the at least one hybridized nucleic acid, thereby providing at least one recombinant nucleic acid that comprises at least one subsequence from each of the at least two parental nucleic acids.

47. The method of claim 46, wherein the codon-varied oligonucleotides are overlapping.

48. The method of claim 46, wherein the at least two parental nucleic acids are less than 90 percent homologous.

49. The method of claim 48, further comprising using the split-pool synthesis format to synthesize the plurality of codon-varied oligonucleotides, wherein the split-pool synthesis format is module-based with a smallest module being a single trinucleotide in length and a larger module being at least 15 nucleotides in length.

50. The method of claim 49, wherein the split-pool synthesis format is a split-pool synthesis format performed in an automated synthesizer.

51. The method of claim 46, wherein the set of fragmented parental nucleic acids is partially produced by cleavage of the two parental nucleic acids with a DNase enzyme.

52. The method of claim 46, wherein at least a portion of the set of fragmented parental nucleic acids is produced by partial chain elongation using a polymerase and one or both of the parental nucleic acids as templates for elongation of one or more hybridized polymerase primer nucleic acids.

53. The method of claim 46, wherein at least a portion of the set of fragmented parental nucleic acids is produced by synthesizing oligonucleotides which correspond to one or more of the at least two parental nucleic acids, which oligonucleotides comprise a plurality of codon-varied oligonucleotides.

54. The method of claim 46, wherein the at least two parental nucleic acids are homologous.

55. The method of claim 46, wherein the at least two parental nucleic acids are non-homologous.

56. The method of claim 46, wherein the hybridizing step comprises hybridizing at least one codon-varied oligonucleotide with at least one additional codon-varied oligonucleotide to provide the at least one hybridized nucleic acid.

57. The method of claim 46, wherein the hybridizing step comprises hybridizing at least one codon-varied oligonucleotide with at least one DNase fragmented parental nucleic acid to provide the at least one hybridized nucleic acid.

58. The method of claim 46, wherein the hybridizing step comprises hybridizing at least one DNase fragmented parental nucleic acid with at least one additional DNase fragmented parental nucleic acid to provide the at least one hybridized nucleic acid.

59. The method of claim 46, wherein the composition is provided by:

aligning homologous nucleic acid sequences to select conserved regions of sequence identity and regions of sequence diversity;

synthesizing a plurality of codon-varied oligonucleotides corresponding to at least one region of sequence diversity;

providing at least one full-length nucleic acid that is identical to, or homologous with, at least one of the homologous nucleic acid sequences;

fragmenting the at least one full-length nucleic acid by DNase cleavage; and,

mixing the resulting set of nucleic acid fragments with the plurality of codon-varied oligonucleotides, thereby providing the composition comprising at least one set of fragmented parental nucleic acids corresponding to the at least two parental nucleic acids, the set of fragmented parental nucleic acids comprising a plurality of codon-varied oligonucleotides.

60. The method of claim 59, wherein the method of recombining at least two parental nucleic acids to provide at least one recombinant nucleic acid is performed in an integrated system comprising a computer or computer readable medium comprising a data set corresponding to a set of character strings corresponding to the aligned homologous nucleic acid sequences and a sequence comparison instruction set for optimizing alignment of the aligned homologous nucleic acid sequences.

61. The method of claim 59, the synthesizing step comprising codon-based coupling chemistry.

62. The method of claim 46, wherein the composition is provided by:

selecting non-homologous nucleic acids to be recombined,

synthesizing a plurality of codon-varied oligonucleotides corresponding to at least one subsequence from each of the at least two parental nucleic acids;

providing at least one full-length nucleic acid that is identical to, or homologous with, at least one of the parental nucleic acid sequences;

fragmenting the at least one full-length nucleic acid by DNase cleavage; and,

mixing the resulting set of nucleic acid fragments with the plurality of codon-varied oligonucleotides, thereby providing the composition comprising at least one set of fragmented parental nucleic acids corresponding to the at least two parental nucleic acids, the set of fragmented parental nucleic acids comprising a plurality of codon-varied oligonucleotides.

63. A method of recombining homologous or non-homologous nucleic acid sequences having low sequence similarity, the method comprising:

recombining at least one set of fragmented nucleic acids with a set of cross-over codon-varied oligonucleotides, which oligonucleotides individually comprise a plurality of sequence diversity domains corresponding to a plurality of sequence diversity domains from homologous or non-homologous nucleic acids with low sequence similarity, thereby producing a recombinant nucleic acid.

64. The method of claim 63, further comprising selecting the recombinant nucleic acid for at least one desired trait or property.

65. The method of claim 63, the method further comprising fragmenting at least one of the homologous or non-homologous nucleic acids to provide the set of fragmented nucleic acids.

66. The method of claim 65, wherein the at least one homologous or non-homologous nucleic acid is fragmented with a DNase enzyme.

67. The method of claim 63, the method further comprising synthesizing a plurality of oligonucleotide fragments corresponding to at least one homologous or non-homologous nucleic acid, thereby providing the at least one set of fragmented nucleic acids.

68. The method of claim 67, wherein the at least one nucleic acid is less than 90 percent homologous.

69. The method of claim 68, wherein the plurality of oligonucleotide fragments is synthesized using the split-pool synthesis format, wherein the split-pool synthesis format is module-based with a smallest module being a single trinucleotide in length and a larger module being at least 15 nucleotides in length.

70. The method of claim 69, wherein the split-pool synthesis format is a split-pool synthesis format performed in an automated synthsizer.

71. A method of recombining a plurality of parental nucleic acids, the method comprising: ligating a set of a plurality of codon-varied oligonucleotides, the set comprising a plurality of nucleic acid sequences corresponding to a plurality of the parental nucleic acids to produce at least one recombinant nucleic acid encoding a full-length protein.

72. The method of claim 71, wherein the plurality of parental nucleic acids is less than 90 percent homologous.

73. The method of claim 72, wherein the plurality of codon-varied oligonucleotides is synthesized using the split-pool synthesis format, wherein the split-pool synthesis format is module-based with a smallest module being a single trinucleotide in length and a larger module being at least 15 nucleotides in length.

74. The method of claim 73, wherein the split-pool synthesis format is a split-pool synthesis format performed in an automated synthesizer.

75. The method of claim 71, the set comprising at least a first oligonucleotide which is complementary to at least a first of the parental nucleic acids at a first region of sequence diversity and at least a second oligonucleotide which is complementary to at least a second of the parental nucleic acids at a second region of diversity.

76. The method of claim 71, further comprising selecting the at least one recombinant nucleic acid encoding a full-length protein for at least one desired trait or property.

77. The method of claim 71, wherein the set of a plurality of oligonucleotides is ligated with a ligase.

78. The method of claim 71, wherein the set of a plurality of oligonucleotides is hybridized to a first parental nucleic acid and ligated with a ligase.

79. The method of claim 71, wherein the plurality of parental nucleic acids are homologous.

80. The method of claim 79, wherein the method of recombining a plurality of parental nucleic acids is performed in an integrated system comprising a computer or computer readable medium comprising a data set corresponding to a set of character strings corresponding to the plurality of homologous parental nucleic acids and a sequence comparison instruction set for optimizing alignment of the plurality of homologous nucleic acids.

81. The method of claim 71, wherein the set of a plurality of oligonucleotides comprises a set of overlapping codon-varied oligonucleotides.

82. The method of claim 71, the method further comprising hybridizing the set of a plurality of codon-varied oligonucleotides to at least one of the plurality of parental nucleic acids, elongating the oligonucleotides with a polymerase and ligating the resulting elongated oligonucleotides to produce a nucleic acid encoding a substantially full-length protein.

83. The method of claim 82, further comprising selecting the nucleic acid encoding a substantially full-length protein for at least one desired trait or property.

84. A composition comprising a library of codon-varied oligonucleotides comprising a plurality of codon-varied oligonucleotide member types, the oligonucleotide member types corresponding to a plurality of subsequence

regions of a plurality of members of a selected set of a plurality of homologous or non-homologous target sequences, which member types comprise a plurality of members with at least one region of similarity and at least one region of diversity, the region of diversity comprising at least one codon difference.

85. The composition of claim 84, wherein the plurality of oligonucleotide member types are present in non-equimolar amounts.

86. The composition of claim 84, the plurality of subsequence regions comprising a plurality of non-overlapping sequence regions of the selected set of a plurality of target sequences, wherein the target sequences are homologous.

87. The composition of claim 84, wherein the plurality of oligonucleotide member types each have a sequence identical to at least one subsequence from at least one of the selected set of target sequences, wherein the target sequences are identical.

88. The composition of claim 84, wherein the plurality of oligonucleotide member types comprise a plurality of homologous oligonucleotides corresponding to a homologous region from the selected set of a plurality of homologous target sequences, wherein each of the plurality of homologous oligonucleotides comprise at least one codon varied subsequence.

89. The composition of claim 84, further comprising at least one of: a polymerase, a thermostable DNA polymerase, a ligase, a thermostable DNA ligase, a nucleic acid synthesis reagent, a buffer, a salt, magnesium, and at least one nucleic acid sequence comprising at least one of the plurality of members of the selected set of homologous target sequences.

90. The composition of claim 84, wherein the plurality of oligonucleotide member types is selected by aligning the plurality of homologous target sequences, determining at least one region of identity and at least one region of diversity and synthesizing the oligonucleotides to encode at least a portion of the at least one region of identity, or at least a portion of the at least one region of diversity,

or at least a portion of both the at least one region of identity and at least one region of diversity.

91. The composition of claim 84, wherein the plurality of oligonucleotide member types comprise at least one member type comprising at least one sequence diversity domain.

92. The composition of claim 84, wherein the plurality of oligonucleotide member types comprise a plurality of sequence diversity domains.

93. The composition of claim 92, wherein the plurality of sequence diversity domains correspond to adjacent sequence regions on a plurality of the plurality of homologous nucleic acids when the homologous nucleic acids are aligned.

94. The composition of claim 84, wherein the library comprises a set of cross-over codon-varied oligonucleotides, each oligonucleotide member of the set of cross-over codon-varied oligonucleotides comprising a plurality of sequence diversity domains corresponding to a plurality of homologous nucleic acids.

95. An integrated system comprising a computer or computer readable medium comprising a data set corresponding to a set of character strings corresponding to a set of codon-varied oligonucleotides.

96. The integrated system of claim 95, wherein the codon-varied oligonucleotides are overlapping.

97. The integrated system of claim 95, wherein the system further comprises a sequence comparison instruction set for optimizing alignment of homologous nucleic acid sequences.

98. The integrated system of claim 95, wherein the system further comprises an automatic synthesizer coupled to an output of the computer or computer readable medium, which automatic synthesizer accepts instructions from the computer or computer readable medium, which instructions direct synthesis of the set of codon-varied oligonucleotides.

99. The integrated system of claim 98, further comprising one or more robotic control elements for incubating, denaturing, hybridizing, and elongating the set of overlapping codon-varied oligonucleotides.

100. The integrated system of claim 99, further comprising a detector for detecting a nucleic acid produced by elongation of the set of codon-varied oligonucleotides, or an encoded product thereof.

1/1

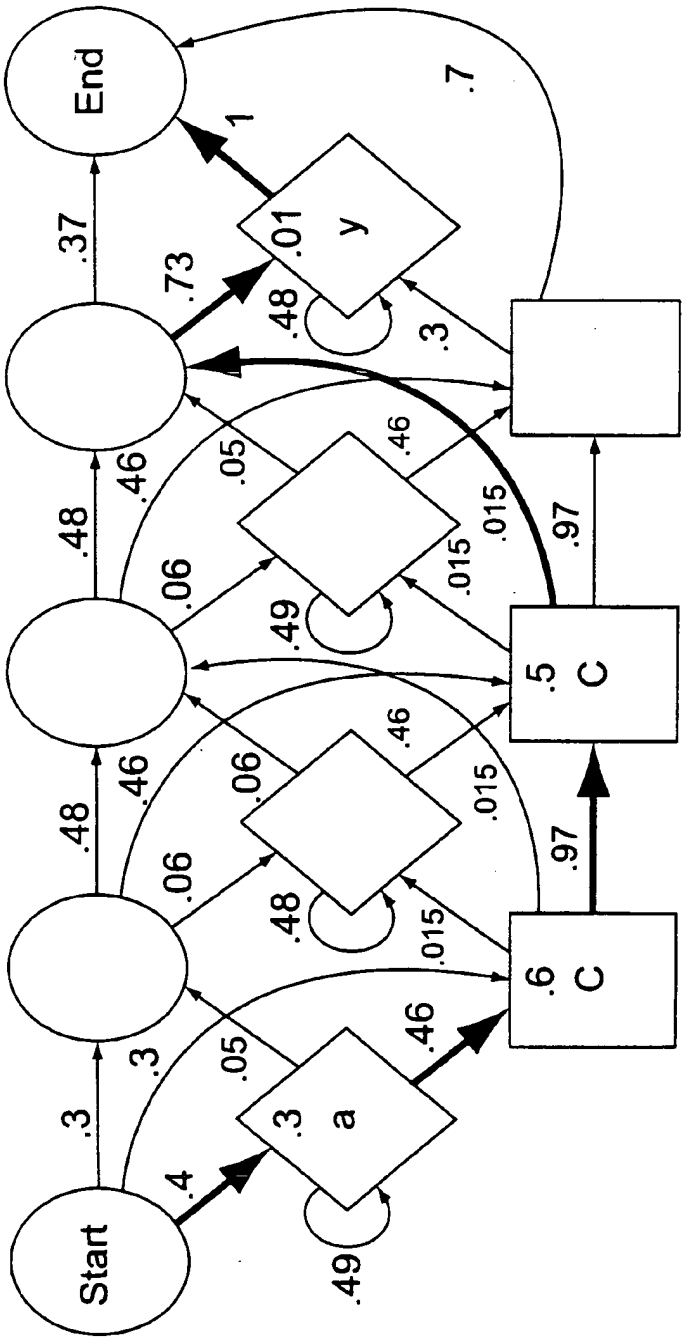

End
Start
y
a
C
C
.37
.73
.01
.48
.3
.7
1
.48
.46
.05
.46
.06
.49
.015
.015
.97
.48
.46
.06
.46
.06
.48
.015
.015
.97
.3
.3
.05
.46
.4
.49
.3
.5
.6

Fig. 1